Facultad de ciencias económicas y empresariales

BIG DATA E INTELIGENCIA ARTIFICIAL EN LA PREDICCIÓN Y MEJORA DE LOS PROCESOS DE VENTA

Autor: Ignacio Peña Sanz

Director: Raúl González Fabre

MADRID | Abril 2024

ÍNDICE

ÍNDICE DE FIGURAS

RESUMEN

Este trabajo de investigación versa sobre las distintas aplicaciones de Big Data e Inteligencia Artificial que utilizan las empresas para mejorar su proceso de ventas, adelantarse a ciclos económicos, mejorar sus estrategias de marketing y, en definitiva, obtener una mejor posición frente a sus competidores.
Tras introducir la problemática actual, se expondrán los conceptos de Big Data e Inteligencia Artificial, sus características y su potencial en el mundo empresarial. También se realizará una exposición de las posibles opciones que tienen las empresas para utilizar a modo de estrategia en el proceso de venta, haciendo un recorrido desde los algoritmos de predicción, los softwares de CRM y los sistemas de recomendación.

En la actualidad, son miles de datos los que se generan al realizarse una transacción, los cuales resultan ser muy valiosos para las empresas que quieren entender el comportamiento de los consumidores, así como mejorar distintos aspectos de sus procesos de venta. Entre estas empresas se encuentra Walmart, una de las empresas pioneras en el análisis de datos y que hoy en día dedica una gran parte de sus ingresos a seguir mejorando estas tecnologías. Por ello, se utilizará Walmart como caso de estudio para ilustrar algunos de los algoritmos de predicción expuestos teóricamente y de esta forma poder compararlos y analizarlos en cuanto a su efectividad.
Además, se expondrá y aplicarán diversos algoritmos que puedan mejorar el modelo como es el caso de Deep Learning u Optimización Hiperparamétrica, así como de Inteligencia Artificial Explicativa que ayuden a la interpretación del modelo de predicción de creado.
Finalmente se presentarán unas conclusiones sobre los resultados obtenidos y de la investigación realizada.

Palabras clave

Big Data, Inteligencia Artificial, Ventas, Algoritmos de predicción, Inteligencia Artificial Explicativa.

ABSTRACT

This research work discusses the various applications of Big Data and Artificial Intelligence that companies use to improve their sales process, get ahead of economic cycles, enhance their marketing strategies, and ultimately achieve a better position against their competitors. After introducing the current issues, Big Data and Artificial Intelligence concepts will be outlined, along with their characteristics, and potential in the business world. It will also be presented the possible options companies have for using these as a strategy in the sales process, going through prediction algorithms, CRM software, and recommendation systems.

Currently, thousands of data points are generated when a transaction takes place, which are very valuable for companies wanting to understand consumer behavior, as well as improve different aspects of their sales processes. Among these companies is Walmart, a pioneer in data analysis, which today dedicates a large part of its income to continue improving these technologies. Therefore, Walmart will be used as a case study to illustrate some of the theoretically exposed prediction algorithms, to compare them and analyze their effectiveness.

Additionally, various algorithms that can improve the model will be presented and applied, such as Deep Learning or Hyperparameter Optimization, as well as Explainable Artificial Intelligence that aids in the interpretation of the created predictive model.
Finally, conclusions will be presented about the results obtained and the research carried out.

Keywords

Big Data, Artificial Intelligence, Sales, Predictive algorithms, Explainable Artificial Intelligence.

1. INTRODUCCIÓN

1.1. OBJETIVOS

El objetivo principal de este Trabajo es analizar las ventajas que tiene el uso del Big Data e Inteligencia Artificial en los procesos de ventas de las empresas.
Objetivos secundarios:

- Conocer el estado del arte de las técnicas de Big Data e Inteligencia Artificial utilizadas en marcos empresariales.
- Evaluar las distintas formas de optimizar el proceso de ventas.
- Aplicar los algoritmos de predicción a la base de datos de ventas de la empresa propuesta (Walmart) y comparar la respuesta de cada uno, determinando el óptimo.
- Investigar sobre técnicas más avanzadas de IA que mejoren el modelo óptimo y ayuden a explicar los resultados obtenidos.

1.2. METODOLOGÍA

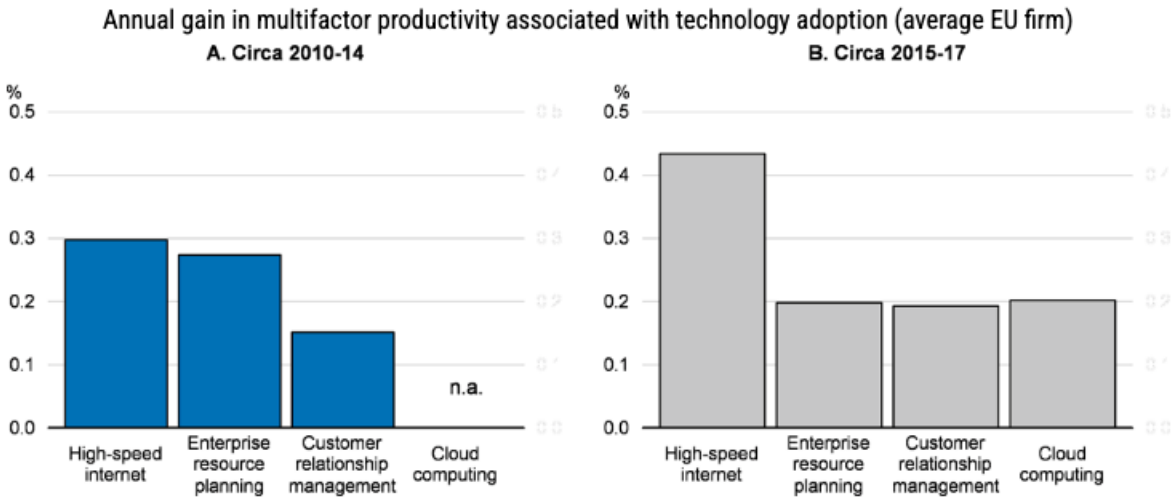
La metodología del trabajo consistirá en una revisión de artículos académicos y técnicos sobre la utilización del Big Data e Inteligencia Artificial en entornos empresariales con un enfoque fundamentalmente teórico. De la misma forma se expondrá un marco tecnológico desde un punto de vista teórico sobre las herramientas de predicción y regresión, softwares de análisis de datos y algoritmos de recomendación, todos ellos bajo el contexto de ser herramientas utilizadas durante el proceso de venta.
A continuación, basándome en los informes y datos sobre la empresa Walmart obtenidos tanto de su página web como de otras páginas de contenidos empresariales, expondré el caso de estudio donde se analizará las herramientas que utiliza esta empresa.
Siguiendo con un enfoque más práctico, llevaré a cabo un análisis de los algoritmos de predicción estudiados durante las asignaturas de Machine Learning cursadas en la Universidad y los aplicaré sobre una base de datos obtenida de la página web Kaggle. Utilizaré esta base de datos para poner a prueba los algoritmos de predicción estudiados y comprobar cuál de ellos resulta más efectivo. Se expondrán las posibles aplicaciones de la Inteligencia Artificial sobre

los resultados de las predicciones arrojados por los algoritmos, de forma que se pueda estudiar su complementariedad. Para finalizar, se realizará una reflexión sobre el algoritmo óptimo y el margen de mejora que tiene, analizando posibles algoritmos avanzados que pudieran serle de aplicación.

1.3. CONTEXTUALIZACIÓN DEL PROBLEMA

Es evidente que, en la actualidad, las empresas que no utilizan análisis de datos se encuentran en plena desventaja con respecto a las que sí lo hacen. Ha sido la propia Organización para la Cooperación y el Desarrollo Económicos, la que ha querido detallar el alcance actual de la utilización del Big Data e Inteligencia Artificial (OECD, 2019, p. 9). En concreto, el análisis de datos y las herramientas digitales pueden *"ayudar en el incremento de la productividad, desde el diseño del producto y los procesos productivos, la automatización de tareas rutinarias, desarrollar ciertas tareas de forma remota y la mejora de los clientes o proveedores con la empresa, entre otras cosas"*.

Figura 1. **La adopción de tecnologías digitales provoca el aumento de la productividad de la compañía**



Note: The exact sub-periods (based on available data) are the following: 2011-14 and 2014-17 for high-speed internet, 2009-13 and 2013-17 for Enterprise Resource Planning software, 2009-14 and 2014-17 for Customer Relationship Management software, and 2014-16 for cloud computing. The effects correspond to the average annual change in the adoption of each technology among EU firms with at least ten employees, multiplied by the elasticity between digital adoption and multifactor productivity growth estimated by Gal et al. (2019). To avoid potential double counting due to collinearity between the adoption of different technologies, the effects are computed using the contribution of each technology to the first principal component of the adoption rate of the selected technologies, and the sensitivity of productivity growth to this first principal component (Table 2, column 7 in Gal et al., 2019). The effects presented correspond to productivity gains after three years, which are obtained by iterating on the error-correction model on which the estimation relies.

Source: OECD calculations based on Gal, P., G. Nicoletti, T. Renault, S. Sorbe and C. Timiliotis (2019), "Digitalisation and Productivity: In Search of the Holy Grail – Firm-Level Empirical Evidence from EU Countries", *OECD Economics Department Working Papers*, No. 1533, OECD Publishing, Paris.

Fuente: OECD (2019), p.11

De acuerdo con Joshi, Spilbergs y Mikelsone (2023), la introducción de estas tecnologías ha derivado en la "Industria 4.0", es decir, la cuarta revolución industrial.

Volviendo al informe sobre Digitalización en la empresa de la OECD, es importante resaltar algunos datos importantes. Por ejemplo, la Organización ha estimado que, en las empresas que utilizan computación en la nube, el aumento del 10% en la proporción de las empresas que utilizan dicha tecnología se relaciona con un incremento del 1.4% en la productividad de una empresa promedio de la industria, después de un año de su implementación. (OECD 2019 p.10)

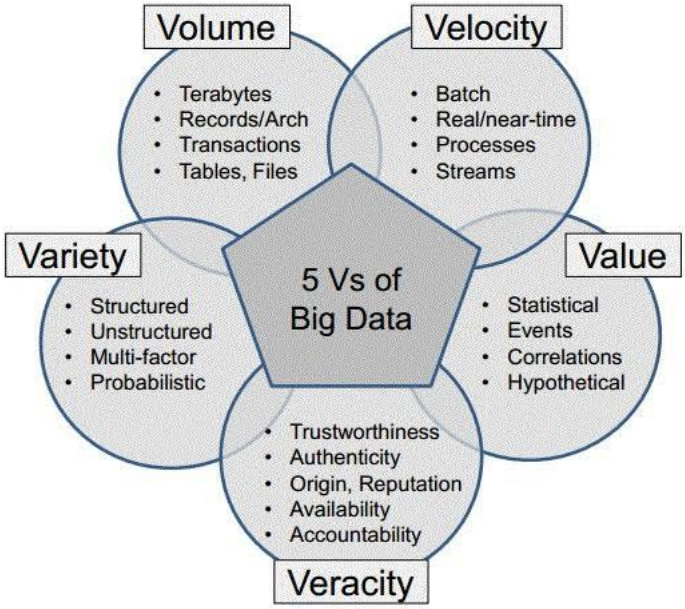
Partiendo de estos datos, y de la premisa de que la digitalización mejora la productividad y eficiencia de las empresas, se van a analizar en concreto las dos tecnologías punteras en la actualidad sobre la que se sustenta el crecimiento en digitalización: el Big Data y la Inteligencia Artificial.

En concreto, nos centraremos en el proceso de ventas y en cómo la tecnología ha afectado al mismo, ya sea al proceso, al producto o a la forma de relacionar a la empresa con los clientes.
La digitalización ha traído consigo la automatización de numerosas tareas, y esto ha producido diversos sentimientos al respecto en los consumidores. Desde algunas que tienen una nefasta consideración en la sociedad (contestadores-robot en llamadas de atención al cliente) hasta otras sin las cuales los consumidores seguramente no elegirían esa empresa.

1.3.1. Big Data y aplicaciones en entornos empresariales

Fue Roger Magoulasin quien formalmente acuñó el término "Big Data" en 2005. Con ello, quiso hacer referencia a grandes cantidades de datos que, al ser tan complejos y grandes, es imposible manejarlos utilizando técnicas de análisis de datos tradicionales (Joshi, Spilbergs, y Mikelsone, 2023). Son innumerables las definiciones que se han hecho sobre el Big Data, las cuales se han ido modificando a medida que se han ido descubriendo nuevas utilidades. En cualquier caso, el Big Data se puede definir como "*una gran colección de datos estructurados, semi-estructurados y sin estructurar, provenientes de diversas fuentes en constante procesamiento que producen a su vez nuevos conocimientos, que no pueden ser almacenados y procesados utilizando las técnicas computacionales tradicionales"* (Ferreiro Piñeiro y Alonso Varona, 2019).
El Big Data, además, se caracteriza por tener cinco dimensiones o características (Moreno, 2014).

Figura 2. **Las 5 V´s del Big Data**



Fuente: López (2017)

En concreto, las cinco dimensiones son:

a) Volumen**.** Se refiere a la cantidad de datos que se recolectan y la necesidad de procesarlos por el valor que aportan los insights que producen. Deben ser cantidades masivas de datos, aunque lo considerado como "alto volumen" puede variar de un sector a otro (Moreno, 2014).

b) Variedad**.** Los datos obtenidos son el resultado combinado de distintas tipologías y fuentes de datos. Todo ello incluye la gestión acerca de los datos estructurados, semi-estructurados y no estructurados. En los últimos años, la aparición de nuevas fuentes que puedan dar lugar a datos, como sensores, dispositivos inteligentes y tecnologías de colaboración social, se presentan en innumerables formas. (Schroeck, Shockley, Smart, Romero-Morales, y Tufano, 2014).

La heterogeneidad en el sentido de la variedad de los datos se puede clasificar alrededor de cuatro ejes, según López (2017):

- Variedad estructural. Se fundamenta en las diferentes formas que existen de representar los datos. Como ejemplo, la forma en la que los datos se aportan a través de un gráfico o de un artículo de investigación.
- Variedad del medio. Es el soporte a través del cual se transmiten los datos. Por ejemplo, un audio y una transcripción.
- Variedad semántica. Es más subjetiva porque depende de antecedentes semánticos creados en la sociedad. Consiste en representar, por ejemplo, la edad a través de números o a través de la imagen de un bebé, adolescente, adulto o anciano.
- Variación y disponibilidad. Los datos pueden ser almacenados y utilizados posteriormente o utilizados en tiempo real. Y a ellos se puede acceder de manera ininterrumpida o intermitente.

c) Velocidad**.** Schroeck, Shockley, Smart, Romero-Morales, y Tufano (2014) lo definen como "datos en movimiento". Con ello, se refieren a la velocidad "*a la que los datos se crean, procesan y analizan"*. La continua mejora de las tecnologías de procesamiento de grandes volúmenes de datos ha propiciado que cada vez esta velocidad sea mayor. Los insights se desarrollan tan rápido que es posible utilizarlos en tiempo real de streaming. Con esto se quiere hacer referencia al concepto de latencia. Esto es, el tiempo que transcurre desde que se obtienen los datos (cualquier tipo desde estructurados a no estructurados) hasta que son presentables, útiles y presentan algún tipo de valor. En esta línea, tal y como establecen Schroeck, Shockley, Smart, Romero-Morales, y Tufano, (2014), para los procesos en los que el tiempo es fundamental (por ejemplo, detección de fraudes o marketing instantáneo), los datos deben analizarse en tiempo real.

d) Veracidad**.** Esta cuarta característica ha sido muy discutida en distintos artículos, ya que inicialmente, en los inicios del uso de la tecnología Big Data, no se utilizaban cantidades tan grandes de datos y el volumen de lo que se genera hoy en día sin duda ha crecido exponencialmente. Por ello, al principio se presuponían todos los datos como verdaderos. Sin embargo, como se ha señalado al inicio de este apartado, se han ido añadiendo características sobre el Big Data a medida que han ido surgiendo nuevos retos y conocimientos sobre el mismo.

Según López (2017), la veracidad hace referencia a la calidad de los datos. La precisión de los datos, la fiabilidad de la fuente y cómo se generaron son solo algunos de los factores que afectan a la veracidad de los mismos.

Hoy en día, la desinformación promovida en parte por las Fake News, es claramente identificable por medio de tecnología Big Data e Inteligencia Artificial. Tal y como indican Moreno Espinosa, Abdulsalam Alsarayreh, y Figuereo Benitez (2024) en su artículo "El Big Data y la inteligencia artificial como soluciones a la desinformación", el Big Data recoge datos en tiempo real y por medio de algoritmos proporcionados por la Inteligencia Artificial es capaz de verificarlos. Esto es altamente beneficioso, ya que, aunque existen empresas verificadoras (FactCheckEU, Iberifier), no son capaces de realizarlo a la misma velocidad que el análisis de datos en tiempo real.

Es por ello por lo que la veracidad de los datos es un pilar y característica esencial del Big Data, utilizando de ejemplo el verificador de Fake News como forma inmediata de ver su utilidad.

e) Valor**.** Consiste en, según BBVA (2021), "*la capacidad de transformar todo ese tsunami de datos en negocio"*. De otra forma, Prometeus Global Solutions (2019) lo define como la utilización de los datos de la forma más rentable y efectiva a la vez. Ser capaz de tomar una inmensa capacidad de datos, aplicar técnicas de Big Data y obtener resultados que vayan a resultar en un beneficio.

En cuanto a las aplicaciones de Big Data en el mundo empresarial, dependerá del sector sobre el que se quiera aplicar. Por ello haremos un repaso, basándonos en el artículo redactado por Joshi, Spilbergs, y Mikelsone (2023), de las mejoras para la empresa que genera el uso de técnicas de análisis de Big Data.

i. Toma de decisiones. El uso de tecnologías Big Data ofrece información de múltiples factores referidos tanto a la empresa de forma interna como a los factores externos. El uso de esta información a la hora de tomar decisiones hace que estas sean más informadas.
ii. Insights de consumidores. A la hora de llevar a cabo estrategias de marketing, por ejemplo, las empresas pueden utilizar datos que muestren el comportamiento del consumidor medio o analizar mejor el target al que dirigen sus productos o servicios.

iii. Eficiencia operacional. Llevándose a cabo un análisis desde la producción hasta las ventas, se pueden encontrar puntos de pérdida de eficiencia.
iv. Tendencias de mercado y análisis competitivo. Conocimiento de las acciones que lleva a cabo la competencia o el propio mercado para poder mantenerse al día de las tendencias.
v. Gestión de riesgos y detección de fraudes. En campos de finanzas o ciberseguridad, en donde el descubrimiento temprano puede ser vital.
vi. Desarrollo de producto e innovación. El análisis Big Data puede proveer información de los clientes y sus demandas y preferencias.

1.3.2. Inteligencia Artificial como herramienta empresarial

La Inteligencia Artificial es un concepto complejo de definir. Aún no existe una definición formal y universalmente aceptada. La Comisión Europea lo define como "*sistemas de software diseñados por humanos que, ante un objetivo complejo, actúan en la dimensión física o digital: a través de la adquisición e interpretación de datos estructurados o no estructurados; y procesando la información derivada de estos datos y decidiendo las mejores acciones para lograr el objetivo dado*" (European Commission, 2022, p. 4). En los últimos años, han proliferado distintos modelos de IA que, gracias al desarrollo de la tecnología computacional, han podido formar sistemas inteligentes que mejoran los análisis de datos y otros tipos de tecnologías implicadas en los procesos de digitalización empresarial.

McCarthy (2007) destacó como uno de los pioneros en explorar la inteligencia artificial, enfocándose en las capacidades necesarias para que estos sistemas emulen el funcionamiento de la mente humana. Los describió como sistemas expertos, analizando qué les faltaba para alcanzar dicho nivel de complejidad. Los sistemas expertos se basan en una serie de conocimientos amplios que son aportados como base e incluyen reglas de lógica e inferencia. Son capaces de, tomando como base el conocimiento proporcionado y las reglas establecidas, tomar decisiones como lo haría un ser humano experto en ese campo. Su eficacia depende de la calidad y profundidad del conocimiento codificado. Esta concepción ya ha sido superada por el Machine Learning, centrado en el aprendizaje a partir de datos. Utiliza algoritmos que identifican patrones y se ajustan de manera automática a través de la experiencia, por lo que no

es necesario actualizar las normas y reglas como ocurre en los sistemas expertos. Por estas razones de mejora en la eficiencia de aprendizaje y de resultados, el Machine Learning superó el uso de los Sistemas Expertos.
En los últimos dos años, el término Inteligencia Artificial ha ganado una gran popularidad debido al surgimiento de los populares chatbots como ChatGPT o Copilot, entre otros. Pero la realidad es que la Inteligencia Artificial lleva desarrollándose desde casi un siglo atrás.
Bill Gates (2023), en una entrevista a La Vanguardia, considera que el avance y desarrollo de la IA es un paso completamente necesario, de igual forma que lo fue el desarrollo de ordenadores personales o el teléfono móvil.

Al fin y al cabo, la Inteligencia Artificial es un paso más en este desarrollo de nuevas tecnologías que ayuda a entretejer la forma en la que se relacionan. Como ejemplo, el desarrollo de tecnologías Iot (Internet of Things) hasta ahora se encontraba más limitado en el sentido de que la extracción de datos en tiempo real y distintas sugerencias que ayudaran en la toma de decisiones empresariales eran más lentas y conllevaban más retraso. Con la implementación de modelos de inteligencia artificial en estos sensores, ya no solo son capaces de extraer información más rápida, sino que la toma de decisiones se puede llevar a cabo incluso de manera autónoma. En el caso del Machine Learning, que es un pilar dentro de la IA, encontramos especializaciones como el Deep Learning, que profundiza en redes neuronales para procesar grandes volúmenes de datos (Big Data). En tanto, la Robótica y la comunicación M2M no son subdivisiones del Machine Learning, sino que se benefician de su aplicación: la Robótica utiliza algoritmos de ML para realizar tareas físicas mediante la combinación de hardware y software, y la tecnología M2M se apoya en estos algoritmos para mejorar la comunicación y optimizar procesos entre dispositivos. Así, se establece una interdependencia entre los sistemas: el hardware ejecuta acciones físicas, los algoritmos de ML proporcionan la inteligencia para guiar estas acciones, y los datos sirven como el combustible que alimenta dichos algoritmos."

Las ventajas de implementar la IA en el entorno empresarial son innumerables. Tal y como establecen Joshi, Spilbergs, y Mikelsone (2023), la integración de la Inteligencia Artificial cataliza el uso de los datos y tecnología y los convierte en valor y competitividad. Si las empresas quieren mantenerse competitivas, no les queda otra opción que la integración de la Inteligencia Artificial en su operativa.

Como se ha desarrollado arriba en relación al Big Data, se exponen la lista de ventajas competitivas que supone la integración de la Inteligencia Artificial en entornos empresariales (Joshi, Spilbergs, y Mikelsone, 2023).

i. Insights guiados por los datos. La inteligencia artificial puede analizar de manera más rápida y efectiva los datos que recibe como inputs y producir insights que ayuden a las empresas en distintas facetas.
ii. Eficiencia. Por ejemplo, a través de la automatización de procesos.
iii. Experiencia de consumidor individualizada. Para conocer las preferencias de los consumidores y ofrecerles, en tiempo real, experiencias que se adapten a dichas preferencias.
iv. Desarrollo de producto e innovación. La IA puede utilizarse para examinar datos de mercado e identificar áreas de innovación o necesidades potenciales. Esto ayuda a las empresas a desarrollar nuevos productos o mejorar los antiguos para mantenerse por delante de la competencia.
v. Optimización de la gestión de la cadena de suministros. La Inteligencia Artificial busca los factores que resten eficiencia en los distintos puntos del proceso y ofrece soluciones tales como la automatización de los mismos.

2. ESTRATEGIAS TECNOLÓGICAS APLICABLES AL PROCESO DE VENTA

2.1. ALGORITMOS DE PREDICCIÓN

Jiménez Estrada y Gómez Herrera (2021) definen los algoritmos de predicción como un conjunto de técnicas que a través del Big Data y la Inteligencia Artificial, buscan realizar una predicción de valores, patrones e informaciones que puedan ser útiles para la empresa en distintas áreas tales como la gestión de la cadena de suministros, el marketing o la experiencia de consumidor personalizada.

Al fin y al cabo, los algoritmos de predicción son aquellos que aprovechan toda la información recibida y la optimizan, de forma que, en lugar de almacenarla y analizarla con técnicas tradicionales de Análisis de Datos, la transforman en datos listos para ser utilizados en diferentes procesos. En relación con el proceso de venta, estos algoritmos pueden ser útiles para conocer la demanda, identificar probabilidades de compra o estimar potenciales clientes. De esta forma se puede llegar a la optimización de cadena de suministros o la personalización de estrategias de marketing.

2.1.1. Tipos de algoritmos de predicción

Los algoritmos de predicción son una técnica de Machine Learning, que se define como "*una rama de la Inteligencia Artificial que se encarga de generar algoritmos que tienen la capacidad de aprender y no tener que reprogramarlos de manera explícita"* (Sandoval, 2018). Es decir, los algoritmos de predicción son aquellos que son entrenados una sola vez (aparte de posibles ajustes que se le puedan ir haciendo) y que, al introducirles una base de datos, operan con ellos de forma que aporta diferentes outputs pivotando sobre una de las modalidades utilizadas, ya sea de aprendizaje supervisado como no supervisado.

Para el entrenamiento de estos algoritmos, existen lo que se llaman dos tipos de aprendizajes. Por un lado, se encuentra el aprendizaje supervisado. En estos modelos, le entregaremos al sistema las características de las preguntas y de las respuestas. De esta forma, el algoritmo aprende cómo serán las futuras preguntas que se le realizarán y cómo deberán ser las respuestas que debe ofrecer.

Dentro de este tipo de aprendizaje, se pueden encontrar algoritmos de clasificación y algoritmos de regresión. Los de clasificación, como su nombre indica, devolverán respuestas sobre si el input que le hemos introducido corresponde o no a cierto grupo de variables. El de regresión, que es el que interesa en este trabajo, obtiene datos pasados y devuelve una serie de valores según la línea de regresión trazada por esos datos. Esta línea determinará la dirección de la predicción que se pretende hacer.

Para ilustrarlo mejor, Sandoval (2018) utiliza el ejemplo del precio de las casas. En ese modelo, se le aporta al algoritmo el precio de distintas casas: pequeñas, grandes, en el campo, en la ciudad, etc. Cuando se le pregunte por el precio de una casa concreta, no ubicada dentro de la base de datos, utilizará las variables que se le han introducido para predecir el valor de la casa en cuestión.

En cuanto al aprendizaje no supervisado, es aquel que agrupa los valores según sus características. Únicamente se le aportan las características, pero no las etiquetas de los valores. Los comparará y unificará en torno a los que considere del mismo grupo.

Otra clasificación posible de los modelos de Machine Learning es la que se realiza en la forma en la que los modelos interaccionan y proporcionan los resultados. En primer lugar, se encuentran los modelos lineales, los cuales buscan trazar una línea entre los valores utilizados de forma que se ajusten todos ellos a la línea y se pueda detectar la presencia de outliers. Otro tipo de modelos son los modelos de árbol, aquellos que utilizan reglas de decisión con cada una de las variables, incluso utilizando reglas no lineales. Dentro de este tipo cabe mencionar los árboles de decisión como tal y los random forest. Por último, las redes neuronales son aquellas que tratan de simular cómo funciona el cerebro humano y replicarlo. Interconectan miles de datos entre sí con resultados y se utilizan en modelos muy complejos lo que resulta en la necesidad de mucho entrenamiento.

Además, cabe mencionar el reinforcement learning. Esta técnica se diferencia del aprendizaje supervisado en que es un agente el que toma decisiones en base a las decisiones que realiza el algoritmo y le aporta feedback sobre las mismas. Cabe apuntar que el agente no es un analista de datos o persona física, sino un algoritmo o parte de un sistema de software diseñado para

aprender a través de la experiencia. No se produce un entrenamiento y prueba de un modelo, si no que sucede a lo largo de la vida del modelo. Necesita una participación del analista con una serie de parámetros sobre los que pueda decidir y clasificar cada decisión como errónea o correcta. Según señala Li (2018), algunas aplicaciones del reinforcement learning son la robótica, como los coches autonómos, herramientas de detección de enfermedades o ciberseguridad, entre otras.

En este TFG, nos centraremos en los modelos de regresión, dado que estos son los que ofrecen datos futuros de los que todavía no se tiene información.

2.2. SOFTWARES DE CRM

El CRM (Customer Relationship Management) se define como "*un conjunto de estrategias de negocio, marketing, comunicación e infraestructuras tecnológicas, diseñadas con el objeto de construir una relación duradera con los clientes, identificando y comprendiendo sus necesidades"* (AEMR, 2002). Esta estrategia de negocio se debe a la amplia variedad de tecnologías que surgen en los últimos años. Previamente, existía lo que se denominaba marketing relacional, una simple estrategia (entiéndase simple como aquella que no conlleva mucha implicación de herramientas tecnológicas) que se centraba en la comunicación de la empresa con los clientes. El CRM es útil en cuanto al proceso de venta ya que registra todas las interacciones que la empresa tiene con el cliente y por tanto guarda datos en cuanto a preferencias y tendencias de compra. Enfoca la optimización del proceso de venta en lo que a la relación con el cliente se refiere.

Si se observa la lista de ventajas empresariales competitivas que se presentaron en el primer apartado de este trabajo, se puede observar que muchas de ellas se basan en el cliente, en observar, analizar y canalizar sus preferencias. Pues bien, el CRM no deja de ser una plataforma en la que volcar toda esa información obtenida a partir del análisis de datos.

Un ejemplo de CRM es la empresa Salesforce. Salesforce es una empresa de computación en la nube (cloud computing) y principalmente SaaS (Software as a Service). Se encarga de proveer softwares de gestión a empresas. Entre ellas está la plataforma de CRM, fuente principal de ingresos (Kotler y Keller, 2015). Las ventajas que conlleva la implementación del

software son, entre otras, la sugerencia de nuevos y potenciales segmentos de clientes, el trabajo en la nube y almacenamiento de datos, gestión de inventarios o comunicación directa con el cliente. De manera creciente, Salesforce provee herramientas que no solo son útiles para la gestión de clientes, sino para la identificación de nuevos segmentos o de potenciales clientes de alta prioridad (Dyer y Gregersen, 2018).

Salesforce utiliza algoritmos de IA y modelos de automatización de tareas que, desde luego, aportan otra visión ya no solo del proceso de ventas, sino de algo que parece más importante en los últimos años, que es mantener al cliente.

2.3. SISTEMAS DE RECOMENDACIÓN

Desde siempre, el "boca a boca" ha sido la mejor forma de publicitar productos o servicios. Los consumidores que adquirían alguno y tenían buenas experiencias, no dudaban en compartirlo tanto con las personas que se encontraban en el momento de tomar la decisión de comprar, como con personas que ni siquiera se habían planteado esa decisión. Con la llegada de internet y los foros de reseñas, esas recomendaciones pasaron a ser escritas y quedar reflejadas en las referencias de productos y servicios. Desde entonces, son innumerables los foros de distintas temáticas en que las personas comparten sus opiniones y experiencias con productos o servicios. Los últimos avances tecnológicos han propiciado una evolución en dichas recomendaciones. Ahora son algoritmos los que, basados en opiniones de otros consumidores o usuarios, las preferencias personales de los compradores, tendencias de mercado y otros factores ofrecen recomendaciones. Incluso muchas veces se ofrecen sin ser solicitadas por el propio consumidor. Esto es lo que denominamos Sistemas de Recomendación.

Los Sistemas de Recomendación se definen como "*técnicas de filtrado de información que nacen con el objetivo de facilitar o asistir al usuario en la toma de una decisión*" (García y Gil, 2020). En otras palabras, son sistemas que aplican una serie de filtros personalizados para el usuario de forma que pueda obtener las recomendaciones en base a factores previamente determinados. Los Sistemas de Recomendación son muy útiles en el proceso de venta ya que son capaces de personalizar la experiencia del usuario y sugerir distintos productos o servicios

basándose en preferencias personales o historial de compra. Son especialmente útiles en el e-commerce.

En la actualidad, son incontables los sitios web que utilizan Sistemas de Recomendación, en especial aquellos dedicados al comercio online. Es propio del ser humano, cuando se encuentra ante una decisión que tomar, buscar recomendaciones, ya sean basadas en experiencias de usuarios o basadas en características de los productos que puedan ajustarse a sus preferencias. Esto es muy útil para el Marketing, ya que pueden mostrarse más productos o con mayor frecuencia para determinados tipos de consumidor o segmentos de población.

El Big Data y la Inteligencia Artificial se integran en el análisis de datos. Los algoritmos de IA se utilizan para extraer, analizar y etiquetar grandes cantidades de datos sobre productos y servicios. Este proceso en simultáneo garantiza que el filtrado de datos se realice de manera efectiva y en línea con los objetivos específicos de obtener información relevante. Son estos algoritmos los que luego realizan el filtrado de forma automática y le proponen al potencial consumidor las mejores referencias.

Siguiendo el artículo de Bron Fonseca y Mar Cornelio (2022), podemos establecer tres componentes dentro de los sistemas de recomendación:

i. Datos de entrada. Es la información que se le aporta al sistema y a partir de la cual el sistema hará el filtrado.
ii. Algoritmo de recomendación. Es el algoritmo seleccionado, que puede ser de distintos tipos como se expondrá después, y combina los datos introducidos con los almacenados previamente para generar las recomendaciones.
iii. Base de datos. Son los datos que contiene el sistema previamente y que le sirven para establecer unos grupos de productos o recomendaciones previas.

De los datos de entrada dependerá el tipo de algoritmo de recomendación que se utilice y, por tanto, los datos de salida. Estos datos de salida pueden ser de una de las dos formas siguientes o ambas en el mejor de los casos:

- Recomendación. Es una lista formada por los objetos más útiles analizados para el usuario.
- Explicación. Es una justificación de cada uno de los objetos de la lista de recomendaciones.

2.3.1. Tipos de algoritmos de recomendación

De nuevo, tal y como exponen en su artículo Bron Fonseca y Mar Cornelio (2022) los tipos de algoritmos de recomendación son:

a. Filtrado colaborativo. Este sistema basa toda la recomendación en la información que posee de los usuarios. Hace un análisis de toda la serie temporal de compras y de preferencias. Se utilizan recomendaciones de productos que le han gustado a usuarios con preferencias y hábitos de consumo similares.
b. Filtrado basado en contenido. En este caso, los algoritmos de recomendación se basan en las características del objeto. Se realiza un análisis de los gustos o intereses del usuario y se le proponen ciertos productos o servicios. Es el que se aplica cuando en tiendas online se muestra la frase "*Te ha gustado este producto, quizá te gusten estos otros"*.
c. Sistemas de recomendación basados en conocimiento. Estos sistemas realizan sugerencias haciendo inferencias y deducciones sobre el usuario. Estas inferencias versan sobre las preferencias personales individuales, no generalizando a todos los demás usuarios. Se distingue de los demás por ser capaz de conocer como un objeto o servicio es capaz de satisfacer las necesidades del usuario, por lo que es capaz de razonar sobre la relación entre la necesidad y la recomendación.
d. Sistemas híbridos. Estos sistemas, como su nombre indica, combinan dos o más sistemas de recomendación para optimizar al máximo la eficacia de las recomendaciones.

2.3.2. Ética de las recomendaciones

Sin embargo, hay cierta parte de la comunidad científica (y de la sociedad) que tiene reservas éticas acerca del uso de estos sistemas en lo que se refiere al uso de datos privados.

En años recientes, la expansión de sistemas de recomendación en dispositivos móviles ha intensificado la percepción de que conversaciones cercanas a estos dispositivos, especialmente si mencionan productos o servicios, pueden influir en los anuncios mostrados posteriormente en el teléfono. Este fenómeno, observado en servicios como Google, parece ocurrir incluso cuando los dispositivos no están conectados a internet, sugiriendo una escucha activa por parte del teléfono (Peco, 2018). Al fin y al cabo, el dispositivo funciona como un recolector de datos. La inercia de la industria de los datos masivos, así como la falta de regulación específica al respecto, hace que los datos, que son, como denominan muchos artículos actualmente (Toonders, 2014) (The Economist, 2017), el nuevo petróleo, se quieran vender al mejor postor. Y para venderlos hay que tener mucha oferta. Se discute acerca de si es ético que un dispositivo personal y privado pueda recolectar información y distribuirla para su venta. Por ejemplo, se pudo demostrar que Google recolecta los datos de ubicación de los teléfonos incluso cuando estos tienen el GPS desactivado (Pastor, 2017).

Otro de los casos que hizo que se empezara a mirar con lupa la recolección de datos fue el caso de Cambridge Analytica (BBC, 2018). Esta empresa se dedicaba al análisis de datos para el diseño de campañas para marcas y políticos. Desarrollaron un test de personalidad que publicaron en Facebook y que completaron más de 265.000 usuarios, quienes autorizaban el acceso a sus perfiles y círculos de amistad. Esto le permitió a la empresa elaborar perfiles psicológicos. Lo que se propuso a hacer fue crear las referidas anteriormente Fake News y orientarlas hacia los perfiles psicológicos en los que fuera a calar el mensaje, por muy falso que fuera.

En el apartado primero, con referencia al Big Data y la Inteligencia Artificial, se expuso como ejemplo la situación en la que pueden tener un papel crucial y beneficioso el mundo de las Fake News. Sin embargo, es posible que, al ser una tecnología con tanto potencial, se utilice de forma perjudicial, utilizando el mismo ejemplo, en la creación y distribución de Fake News. Es ahí donde debe entrar el debate ético y una regulación del uso de los datos.

En concreto, la Unión Europea impulsó la regulación del uso y protección de los datos mediante el Reglamento de Protección de Datos, de 2016, vigente en España desde 2018. Este

Reglamento establece que los usuarios deben dar su consentimiento cuando sus datos se vayan a usar con fines publicitarios, siendo dicho consentimiento plenamente consciente e informado, lo que implica que deben saber a qué fines se van a destinar sus datos. Además, rigen los principios de minimización y limitación, que suponen que los datos recogidos deben ser los estrictamente necesarios y su uso solo puede ser para el que el usuario ha prestado su consentimiento.
Sobre el papel, es una regulación muy rigurosa. Sin embargo, en la práctica, ya existen métodos que consiguen sortear la regulación y recopilar los datos sin consentimiento, como la escucha activa por parte de los dispositivos móviles.

En conclusión, los datos masivos han supuesto un cambio radical en la forma en la que las empresas enfocan sus relaciones con los consumidores, campañas de marketing, producción y gestión de inventarios, etc. Si bien esto ha supuesto un cambio en el paradigma empresarial, hace falta una regulación que oriente y limite la forma en la que se obtienen estos datos.

3. CASO DE ESTUDIO: WALMART

3.1. WALMART

De acuerdo con Britannica (The Editors of Encyclopaedia, 2023), Walmart es una empresa multinacional estadounidense encuadrada dentro del concepto de discount store, esto es, que vende sus productos a un precio menor que el del mercado (Mahmoud y Rasha, 2019). Es una de las compañías más grandes del mundo y conocida por su estrategia empresarial desde su creación en 1962 por Sam Walton. Si bien al principio sólo contaba con una tienda de retail, su estrategia de expansión le llevó a tener más de 24 tiendas abiertas en 1969. Diez años más tarde, al final de la década de 1970, Walmart ya contaba con 276 tiendas operativas en más de once estados de Estados Unidos. Esta rápida expansión se debe a numerosos factores estratégicos que los dirigentes y, en especial, Sam Walton pusieron en marcha. Estos, según Humberto, Ramon, y Emili (2015) son: a) estrategia de precios; b) comunicación con los vendedores; c) inversión en tecnología; d) mejora en los recursos humanos; e) políticas de expansión; f) selección de productos; g) conciencia de costes; y h) atención al cliente.

Dado el tema que nos ocupa, nos centraremos en explicar la política de inversión en tecnología que llevaron a la optimización de la cadena de suministros de Walmart y a uno de los análisis de datos más primitivos.

Ronald Mayer fue el CEO de Walmart desde 1974 hasta 1976, y siempre abogó por el uso de tecnologías digitales que pudieran mejorar la eficiencia de la empresa. Sam Walton adoptó sus ideas, llevando a Walmart a ser una de las compañías que primero adoptó el sistema de Uniform Product Codes (UPC) en los puntos de venta. Este sistema permitía conocer en tiempo real la localización de todos los productos y, por tanto, mejorar la eficiencia en las etapas finales de venta directa al consumidor. Otra de las innovaciones que se introdujeron en ese tiempo fue el reemplazo de las cajas registradoras antiguas por puntos de venta digitales que pudieran hacer un seguimiento del inventario y de las tendencias de los consumidores. Estos datos, posteriormente, se analizarían por los encargados de cada tienda y se repondrían productos en mayor o en menor medida en función de las tendencias recabadas en los puntos de venta. Ello vino a ser un análisis de datos primitivo, dado que no utilizaba ningún tipo de programación más que la capacidad analítica de los trabajadores para descubrir patrones de compras.

Dando un salto a la actualidad, Walmart cuenta con más 20.000 tiendas en 28 países, y tiene una de las nubes de datos privadas más grandes del mundo, con más de 2,5 petabytes de datos cada hora (Marr, 2017). Por esa razón, Walmart ha creado el "Data Café", un centro de análisis de datos en el que, a través de más de 200 canales de datos internos y externos, puede analizar en tiempo real distintas métricas. La empresa considera el análisis de datos fundamental para el éxito de la compañía. En muchas ocasiones, se producen errores humanos o de cálculo en la planificación de, por ejemplo, una campaña de un producto. Sin embargo, estos errores son fácilmente identificables y subsanables bajo el análisis de datos. Una de las anécdotas que expone Marr (2017) para ilustrar el caso es sobre un equipo de ventas de alimentación que no entendían por qué había bajado el volumen de ventas de uno de sus productos que tradicionalmente mejor se habían vendido. Cuando acudieron al Data Café a intentar localizar el origen de dicha bajada, descubrieron que se había producido un problema en el proceso de poner el precio, resultando en el producto más caro de lo que normalmente había estado. Este problema se pudo solucionar en apenas una hora y las ventas volvieron a su volumen normal.
Otro ejemplo es el del análisis en tiempo real de los niveles de venta. Durante Halloween, los analistas pudieron comprobar cómo una galleta que se estaba vendiendo muy bien en todas las tiendas, estaba obteniendo unos resultados pésimos en otras dos. Identificaron el problema y pudieron determinar que la causa era de colocación del stock, solventarlo y recuperar el nivel de ventas esperado. Esto se debe a que el sistema también produce alertas en tiempo real si alguna de las métricas baja de forma considerable.

Los datos de Walmart son obtenidos de una serie muy variada de bases de datos, desde datos relacionados con la meteorología hasta redes sociales o bases de datos de ayuntamientos locales. Esto le genera una cantidad inmensa de datos que analizar, por lo que una gran parte de la labor de los analistas es la discriminación de datos no útiles.

Walmart no sólo mira los datos del pasado o del presente, sino que además intenta predecir cuáles serán los datos del futuro. La compañía busca realizar análisis predictivo de su negocio, diferenciando principalmente dos métodos: online y offline. (Cao, 2021)

En el análisis predictivo online, Walmart utiliza algoritmos de recomendación. Como se ha expuesto anteriormente, Walmart recibe en tiempo real insights sobre el comportamiento de los

consumidores. Pues bien, la empresa utiliza dicha información para mostrar productos que puedan ser de interés a los usuarios que realizan compras a través de su página web. Walmart puede anticiparse a las necesidades y recomendar los productos complementarios, mostrándolos en páginas del carrito de compra o similares.
No sólo eso, sino que también utiliza estas predicciones para decidir la forma en la que disponen los productos en las estanterías de las tiendas físicas, o para decidir qué marcas deben dejar de vender y cuáles impulsar.

En cuanto al análisis predictivo offline, Walmart se centra en intentar estimar la demanda y potenciales ventas de más de 500 millones de productos (Cao, 2021). Se utiliza de forma que se pueda optimizar la cadena de distribución y asegurar que cada comprador vaya a encontrar el producto buscado. Por ejemplo, Walmart utiliza este tipo de predicción para los productos de farmacia. La compañía analiza los datos de compra de productos con y sin prescripción médica. Para los primeros, sabe estimar y predecir la demanda en función de la estación y enfermedades crónicas. Para el segundo, que es más complejo dado que no tiene acceso a los datos de prescripciones médicas por ser confidenciales, hace un análisis de número de ventas de estos tipos de medicamentos, agrupándolos por clases y cruzándolos con los conjuntos de datos que se han mencionado antes: meteorología, ayuntamientos locales, redes sociales, etc. Estos datos cruzados le generan insights a través de los cuales es capaz de conocer la futura demanda (Cao, 2021). Según palabras del Jefe de Análisis de Datos de Walmart, Naveen Peddamail (Marr, 2017), lo que buscan evitar es que los potenciales compradores vayan a Walmart y no encuentren los productos que buscan, pero tampoco que las estanterías se queden llenas de productos que sobran.

3.2. APLICACIÓN PRÁCTICA DE LOS ALGORITMOS DE PREDICCIÓN

En este apartado, procederé a realizar un análisis de la aplicación de los algoritmos de predicción expuestos en el apartado 2 sobre una base de datos que contiene las ventas de Walmart de los años 2010 a 2012. Mediante este estudio se quiere hacer una simulación de las predicciones que realiza la empresa estadounidense y descubrir cuál es la que mejor funciona.

3.2.1. Selección, análisis y preparación de la base de datos

La base de datos ha sido obtenida del sitio web Kaggle, y contiene los datos de venta de 45 tiendas Walmart desde el año 2010 hasta el año 2012 y consta de 6.435 registros. Las variables de la base de datos son las siguientes:

- Store. Indica el código de tienda por el que se identificarán los distintos datos. Van desde el 1 hasta el 45.
- Date. Fecha en la que se producen las ventas.
- Weekly_Sales. Ventas semanales de la tienda.
- Holiday_Flag. Variable binaria que designa si el día señalado es festivo (1) o no (0).
- Temperature. Temperatura del día en que se producen las ventas, en grados Fahrenheit.
- Fuel_Price. Precio de la gasolina en la región para el día señalado.
- CPI. Consumer Price Index, en inglés. Índice de Precios de Consumo, en español.
- Unemployment. Tasa de desempleo en Estados Unidos.

Utilizaremos el lenguaje de programación R para llevar a cabo el análisis predictivo. Una vez cargada la base de datos visualizamos la tabla para comprobar las variables.

Figura 3. **Resumen de las variables de la base de datos**

```
'data.frame':   6435 obs. of  8 variables:
 $ Store       : int  1 1 1 1 1 1 1 1 1 1 ...
 $ Date        : chr  "05-02-2010" "12-02-2010" "19-02-2010" "26-02-2010" ...
 $ Weekly_Sales: num  1643691 1641957 1611968 1409728 1554807 ...
 $ Holiday_Flag: int  0 1 0 0 0 0 0 0 0 0 ...
 $ Temperature : num  42.3 38.5 39.9 46.6 46.5 ...
 $ Fuel_Price  : num  2.57 2.55 2.51 2.56 2.62 ...
 $ CPI         : num  211 211 211 211 211 ...
 $ Unemployment: num  8.11 8.11 8.11 8.11 8.11 ...
```

Lo que se observa en la base de datos es que hay variables que no tienen la categorización correcta. Es el caso de las variables “Date”, “Store” y “Holiday Flag”. Esta última la convertiremos en una variable lógica que indique TRUE cuando el resultado sea sí o (1) y FALSE en el caso contrario. En el caso de la variable “Date” solo habrá que especificar que se trata de fechas. Por último, la variable “Store” tiene la condición de entera, y la convertiremos a character dado que los números del 1 al 45 realmente representan el nombre de cada tienda.

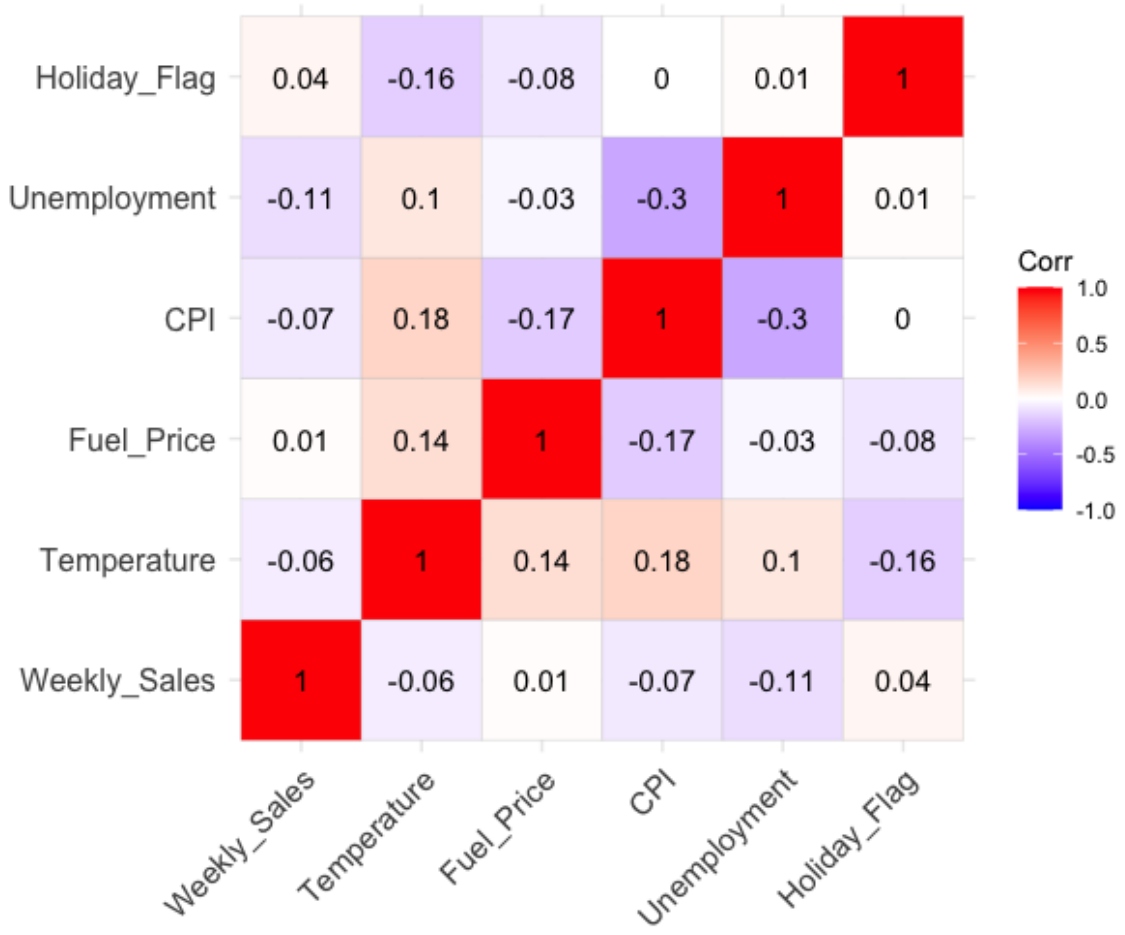
Lo siguiente que haremos será estandarizar la variable “Weekly_Sales”. Como se puede ver en la Figura 2, es la única variable cuyos valores son mucho más altos en comparación al resto. Escalarlas nos permitirá reducir sus valores y otorgarles una media de 0 y desviación típica de 1. Esto ayuda a la interpretación de los coeficientes de regresión y la comparación de cada variable según su contribución al modelo.
Para mejorar la interpretación de los resultados, también se normalizan las demás variables del conjunto de datos.

Figura 4. **Comparación de los valores de la variable "Weekly_Sales" y la misma variable estandarizada**

```
> summary(walmart$Weekly_Sales)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
 209986  553350  960746 1046965 1420159 3818686
> summary(walmart$sales_standar)
       V1
 Min.   :-1.4830
 1st Qu.:-0.8746
 Median :-0.1528
 Mean   : 0.0000
 3rd Qu.: 0.6613
 Max.   : 4.9112
```

El siguiente paso, será producir la matriz de correlación entre variables numéricas. La matriz de correlación es una herramienta que relaciona las variables entre sí y les asigna un valor comprendido entre 0 y 1 según la correlación que haya entre unas y otras. Estos valores pueden llegar a ser negativos si la correlación resulta ser inversa. Un valor de 0 puede indicar que no existe ningún tipo de correlación entre variables.

Figura 5. **Matriz de correlación**



No procede ir variable a variable analizando los resultados, pero sí que considero reseñable comentar dos de estos.

a. CPI y Unemployment. Parece lógico que estas dos variables tengan algún tipo de relación, aunque esta sea negativa. El resultado que arroja la matriz es de -0,3, es decir, una correlación negativa media que implica que a mayor tasa de desempleo menor es el CPI (IPC en español).
b. Temperature y CPI. Esta correlación sí que es más complicada de encontrarle la lógica. Sin embargo, con una puntuación de 0,18, todo parece indicar que hay un mayor CPI (IPC) cuando las temperaturas son más altas. La relación entre ambas variables no es muy fuerte como para ser completamente correlacionales, pero sí que merece la pena resaltarla.

El resto de las variables tienen correlaciones más débiles, es decir, más cercanas a cero. Esto no implica que no estén correlacionadas, dado que incluso puede producirse una correlación no lineal. A pesar de ser un buen análisis preliminar de las variables, no es completamente significativo, ya que se requiere llevar a cabo otros análisis estadísticos. En cualquier caso, para este caso simplemente nos interesa conocer dichas correlaciones para luego poder realizar un mejor análisis de los coeficientes de regresión.

3.2.2. Regresión Lineal

Un modelo de regresión, cómo se ha definido anteriormente, consiste en el análisis de variables que existen dentro de un conjunto de datos y la construcción de una línea recta que modela la relación entre dichas variables. El modelo sigue la siguiente ecuación:

$$Y = a + bX + \epsilon$$

Donde:

- Y es la variable dependiente que se busca predecir. En este caso será Weekly_Sales
- X es la variable independiente utilizada para hacer predicciones. Utilizaremos todas las demás variables, salvo Date, y Store.
- a es el intercepto. Esto es, el valor de Y cuando X es igual a 0.
- b es la pendiente de la línea de regresión.
- ϵ es el término de error.

Se ha decido no incluir la variable Date por el riesgo de multicolineidad que la regresión lineal tiene.

Habiendo realizado en el apartado anterior el análisis de la matriz de correlación de las variables dentro de la base de datos, lo que procede realizar ahora es la partición del modelo en dos conjuntos, el de entrenamiento y el de test o prueba. Para ello, ajustaremos la medida del 80% de los datos para el conjunto de entrenamiento y el 20% de los datos para el de test. Se utilizará una semilla generadora y la función “createDataPartition”. Es importante realizar primero la partición para que la validación de datos funcione correctamente.

Lo siguiente será crear el modelo de regresión sobre el conjunto de entrenamiento. Con ello, se realizará un análisis de los coeficientes que poseen las variables dentro del modelo. Utilizaré la función de R "lm" (linear model). Para el entrenamiento y las predicciones, se utilizará el paquete "Caret" de RStudio.

Figura 6. **Resumen del modelo de regresión lineal construido**

```
Residuals:
    Min      1Q  Median      3Q     Max
-1.7669 -0.8492 -0.2069  0.7004  4.9627

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      -0.009978   0.014302  -0.698  0.48544
Temperature      -0.023882   0.014629  -1.632  0.10265
Fuel_Price       -0.007611   0.014404  -0.528  0.59727
CPI              -0.106591   0.015297  -6.968 3.62e-12 ***
Unemployment     -0.140711   0.014822  -9.493  < 2e-16 ***
Holiday_FlagTRUE  0.148790   0.054461   2.732  0.00632 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9885 on 5145 degrees of freedom
Multiple R-squared:  0.0256,    Adjusted R-squared:  0.02465
F-statistic: 27.03 on 5 and 5145 DF,  p-value: < 2.2e-16
```

Este modelo ha sido construido utilizando todas las variables dependientes y servirá para determinar la relevancia de cada una de ellas dentro del modelo. Para ello nos guiaremos por el p-valor, que mide la probabilidad de obtener los resultados observados cuando la hipótesis nula es verdadera. Esto es, que los resultados sean por pura casualidad. Cuanto más bajo sea, menor probabilidad hay de que se produzcan por casualidad. En base a todo esto, las variables se clasifican en significativas y no significativas. Del primer grupo forman parte las variables CPI y Unemployment, con p-valores extremadamente bajos que indican que tienen un efecto significativo y relevante en las ventas semanales. En este grupo se incluirá también la variable Holiday_Flag, que también cuenta con un p-valor bajo y por tanto un efecto relevante en el

número de ventas. En cuanto a las no significativas, serían Fuel_Price y Temperature, ya que cuentan con un p-valor mucho más elevado que el resto de las variables.

En resumen, las variables más influyentes del modelo son CPI, Unemployment y Holiday_Flag. Esto parece ser bastante lógico dado que las dos primeras son métricas económicas directamente relacionadas con la capacidad adquisitiva del consumidor, quien, realmente, va a realizar las compras. Por su parte, Holiday_Flag, indica los días en los que los clientes van a tener tiempo libre para poder realizar sus compras.

A continuación, procederemos al entrenamiento del modelo. Utilizaremos el método de validación cruzada, en la que el conjunto de datos se divide en partes iguales ("folds") y en cada iteración se utiliza una de las partes como entrenamiento y prueba.

El proceso es relativamente sencillo, utilizando el paquete "Caret", éste realiza la partición de forma interna y genera, a través de inputs que le damos, lo que sería el algoritmo de regresión lineal.
Tras generar las predicciones basándolas en el algoritmo, obtenemos los resultados del modelo creado.

Figura 7. **Rendimiento del modelo de regresión lineal**

RMSE	Rsquared	MAE
0.98417560	0.02441607	0.82744372

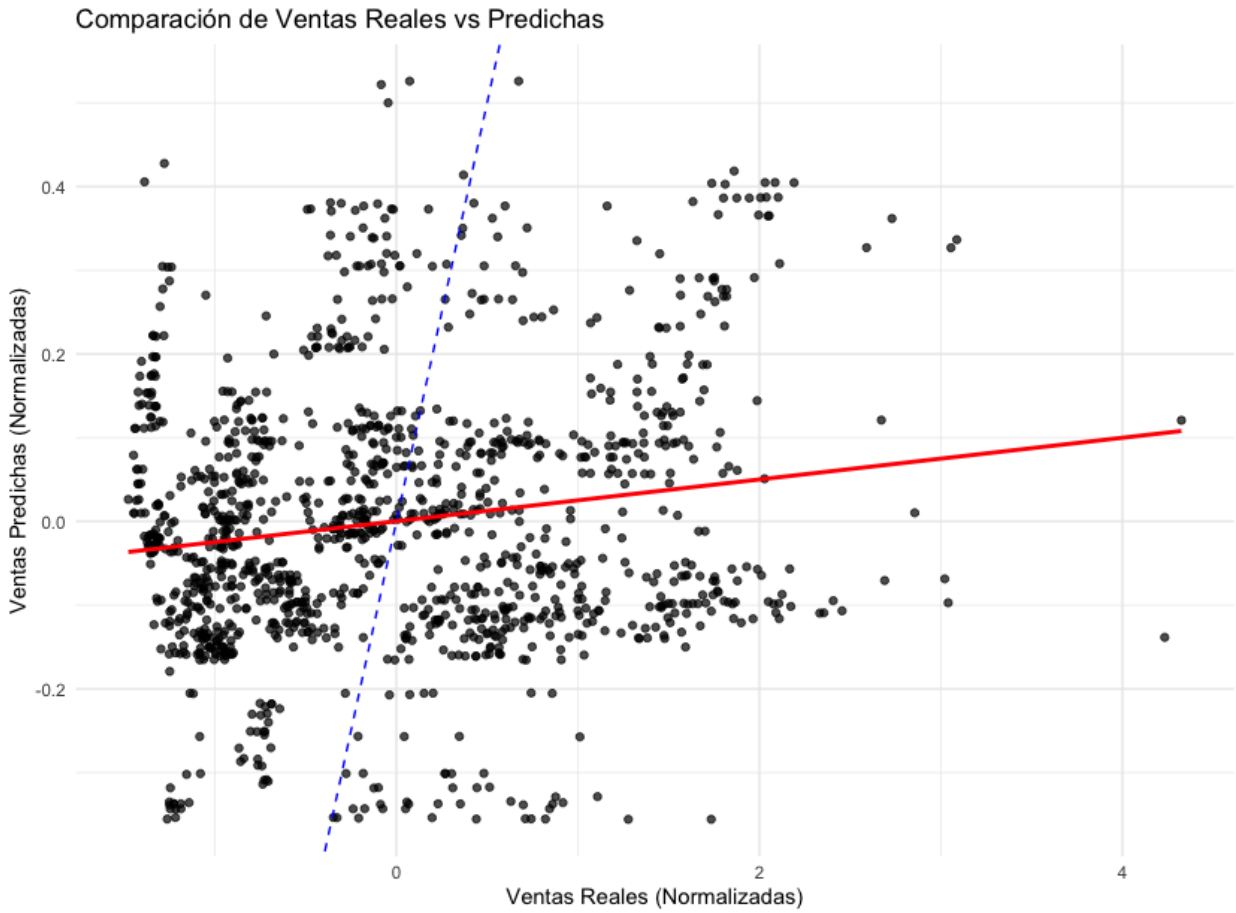
Analizándolo por partes:

- RMSE. Indica el Error Cuadrático Medio. Mide la cantidad de error que hay en las predicciones. En este caso, al ser casi igual que la desviación estándar, indica que el modelo puede no haber capturado bien la variabilidad de los datos.
- Rsquared. Es el Coeficiente de Determinación. Es el porcentaje de variable que se consigue explicar a través del modelo. En este caso es del 2,44%, lo que implica que hay mucha variabilidad que no ha sido posible explicar a través de la predicción. Hay muchos factores que influyen en las ventas y que no han sido contemplados en el modelo.

- MAE. Es el Error Absoluto Medio. Indica la cantidad promedio de error en las predicciones. Al ser de 0,82 en la escala de datos normalizados indica una precisión moderada del modelo.

Generando un gráfico que muestre la diferencia entre los valores predichos y los reales se podrá observar el ajuste del modelo.

Figura 8. **Comparación de los valores predichos y los valores reales**



Para que el modelo sea acertado, debe alinear los puntos en torno a la línea roja. Como se puede apreciar en este caso, hay demasiados valores muy alejados de la recta, lo que induce a pensar que el modelo tiene demasiado sobreajuste o que hay variables que no se han podido explicar con el algoritmo. El modelo tiene un poder predictivo limitado.

Por ello, es una buena opción explorar otras alternativas. Estas alternativas pueden ser realizadas de forma independiente o incorporadas al modelo que ya se ha construido de regresión lineal mediante ENSEMBLES. Estos son una técnica de aprendizaje automático que combina las predicciones de distintos modelos para mejorar la calidad del modelo final.

3.2.3. KNN

El algoritmo de K-Nearest Neighbors (KNN, por sus siglas) es un algoritmo de aprendizaje supervisado que se utiliza tanto para clasificación como para regresión. La teoría detrás del algoritmo presupone que los datos cercanos entre sí en el espacio de características son más propensos a tener etiquetas similares.

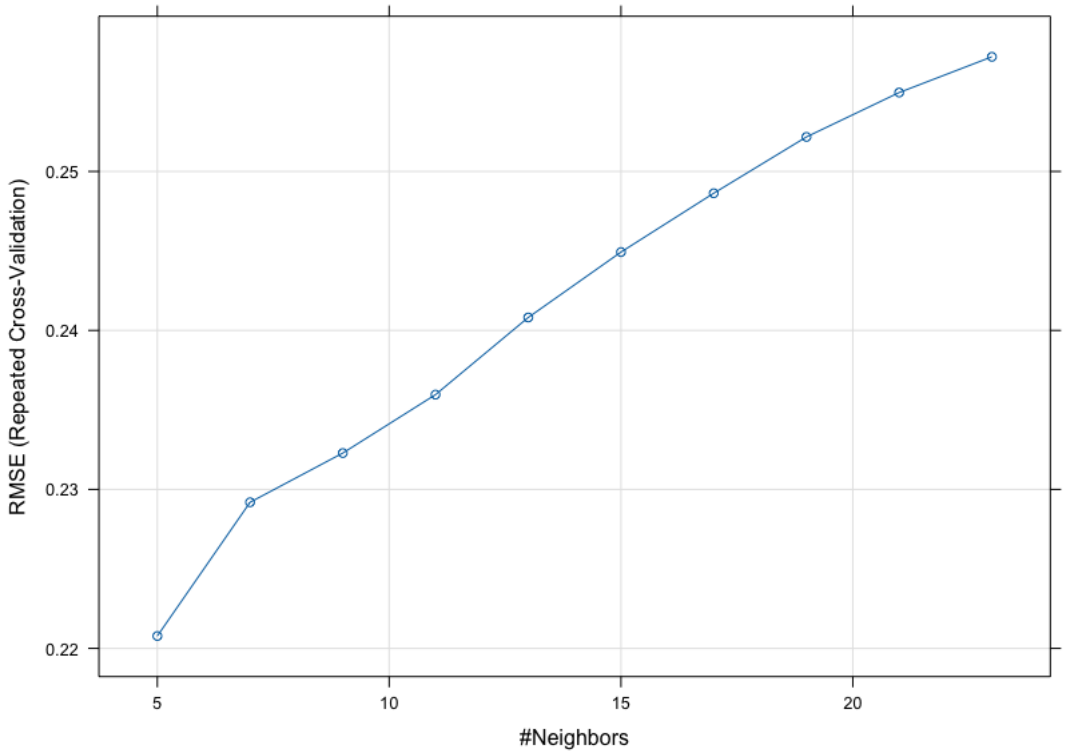
La K en el algoritmo KNN representa el número de vecinos cercanos que se considerarán para tomar la decisión. El algoritmo calcula el promedio de los valores de K de los vecinos más cercanos. Por ello, es muy importante determinar cuál es el número óptimo de K.
Entrenando el modelo, se genera una lista con los posibles valores de K y sus correspondientes valores de rendimiento. De igual forma que en el modelo de regresión lineal, utilizaremos todas las variables independientes salvo Date y Store. En este caso, no se incluye la variable Date debido a que podría interferir en la escala que KNN utiliza para realizar la distancia entre los K vecinos más cercanos. Por tanto, la lista de variables queda de la siguiente forma:

- Variable dependiente: Weekly_Sales
- Variables independientes: CPI, Unemployment, Holiday_Flag, Fuel_Price y Temperature

Figura 9. **Rendimiento del conjunto de entrenamiento de KNN**

k	RMSE	Rsquared	MAE
5	0.2207763	0.9518977	0.1170618
7	0.2291894	0.9481582	0.1210286
9	0.2322850	0.9468302	0.1228845
11	0.2359619	0.9453191	0.1242907
13	0.2408140	0.9432214	0.1261809
15	0.2449276	0.9413550	0.1282407
17	0.2486292	0.9396415	0.1299774
19	0.2521722	0.9379562	0.1313426
21	0.2549639	0.9365803	0.1324752
23	0.2572117	0.9354139	0.1333307

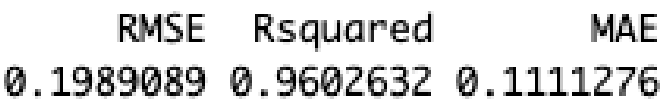
Figura 10. **Valor de K y RMSE correspondiente**



A la vista de los datos mostrados en las figuras 8 y 9, elegimos K=5 como valor óptimo, debido a que es el valor con menor RMSE y mayor RSquared de todos. Recordando la definición de estos dos conceptos, K=5 dará lugar a un modelo en el que la cantidad de error sea la mínima posible y explicará mayor variabilidad.

Tras entrenar al modelo se realizan las predicciones sobre el conjunto de prueba, obtenemos los siguientes indicadores de rendimiento del modelo.

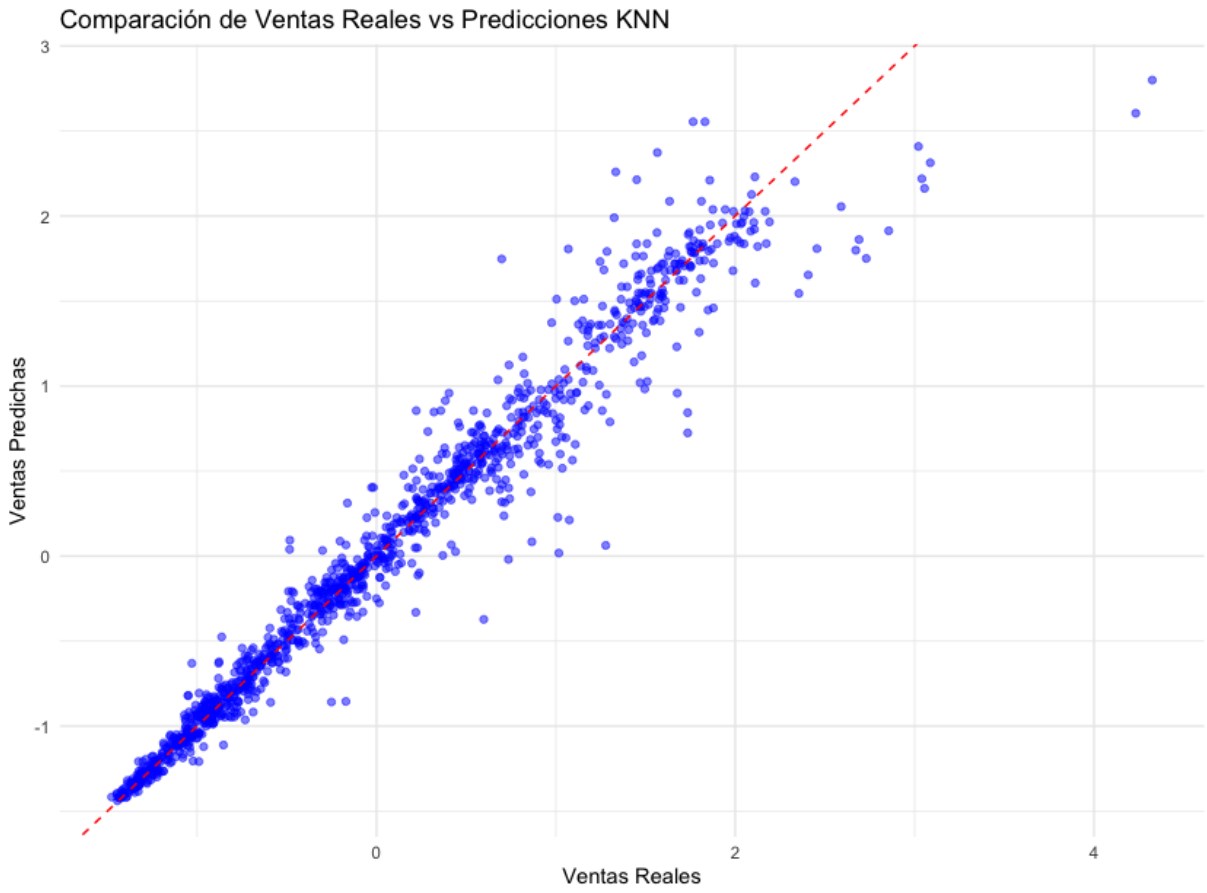
Figura 11. **Rendimiento del modelo KNN**



- RMSE. Tiene un valor bastante bajo, lo que indica que las predicciones están bastante cerca de los valores reales.
- Rsquared. Es extremadamente alto, por lo que el modelo explica el 96% de la variabilidad en los datos de las ventas. El modelo se ajusta muy bien a los datos históricos.
- MAE. Igual que el RMSE, demuestra los errores, pero sin elevarlos al cuadrado. Tiene un valor bastante bajo, lo que es buena señal.

Se procede a representar gráficamente las predicciones.

Figura 12. **Comparación de los valores predichos y los valores reales (KNN)**



En comparación con el modelo de regresión lineal, se puede observar que este modelo ajusta mucho mejor las predicciones en torno a la línea de predicción. El hecho de que estos puntos se encuentren en esa posición implica que las predicciones son bastante precisas. Al extenderse, además, a lo largo de toda la gráfica, se aprecia como el modelo captura la tendencia en los datos.

En general, el modelo es bueno y es capaz de realizar predicciones muy precisas.

3.2.4. Random Forest

Random Forest es un algoritmo de aprendizaje automático utilizado para problemas de clasificación y regresión. Funciona construyendo múltiples árboles de decisión durante el entrenamiento y emitiendo la predicción media de los árboles individuales. Se entrena un árbol de decisión en cada uno de los subconjuntos generados con los datos de entrenamiento con

reemplazo y crecen hasta sus longitudes máximas. Cada vez que se considera una división, a diferencia de los árboles de decisión, se selecciona un subconjunto aleatorio de datos. La predicción final es el promedio de las salidas de todos los árboles.

Se procede a entrenar el modelo de Random Forest y evaluar la cantidad óptima de variables que debe utilizar. En este caso, utilizaremos todas las variables independientes salvo Store. A diferencia de los modelos de regresión lineal y KNN, en este caso sí que se incorpora la variable Date, ya que Random Forest captura bien las características de la fecha y no aparecen problemas de multicolineidad. Por tanto, la lista de variables queda de la siguiente forma:

- Variable dependiente: Weekly_Sales
- Variables independientes: CPI, Unemployment, Holiday_Flag, Fuel_Price, Temperature y Date

Figura 13. **Rendimiento del modelo de entrenamiento Random Forest**

mtry	RMSE	Rsquared	MAE
2	0.46853857	0.9246185	0.349658106
26	0.04208223	0.9980485	0.009604457
51	0.00729103	0.9999305	0.001057169

Siguiendo la misma línea analítica que se realizó con los anteriores modelos, desglosamos la información según las métricas. Es importante puntualizar que “mtry” hace referencia al número de variables que se han probado dentro del modelo. El propio modelo nos indica que el número óptimo de variables es 51, por las siguientes razones:

- RMSE. En este caso, es un valor extremadamente bajo, lo que indica que no existe apenas diferencia entre los valores reales y los predichos por el modelo. Este valor indica una altísima precisión.
- Rsquared. Al contrario que el RMSE, es un valor extremadamente alto, casi 1, lo que indica que es capaz de recoger toda la variabilidad del modelo y explicar toda la varianza de la variable dependiente.
- MAE. De forma similar al RMSE, al ser tan bajo refleja una precisión muy elevada en las predicciones.

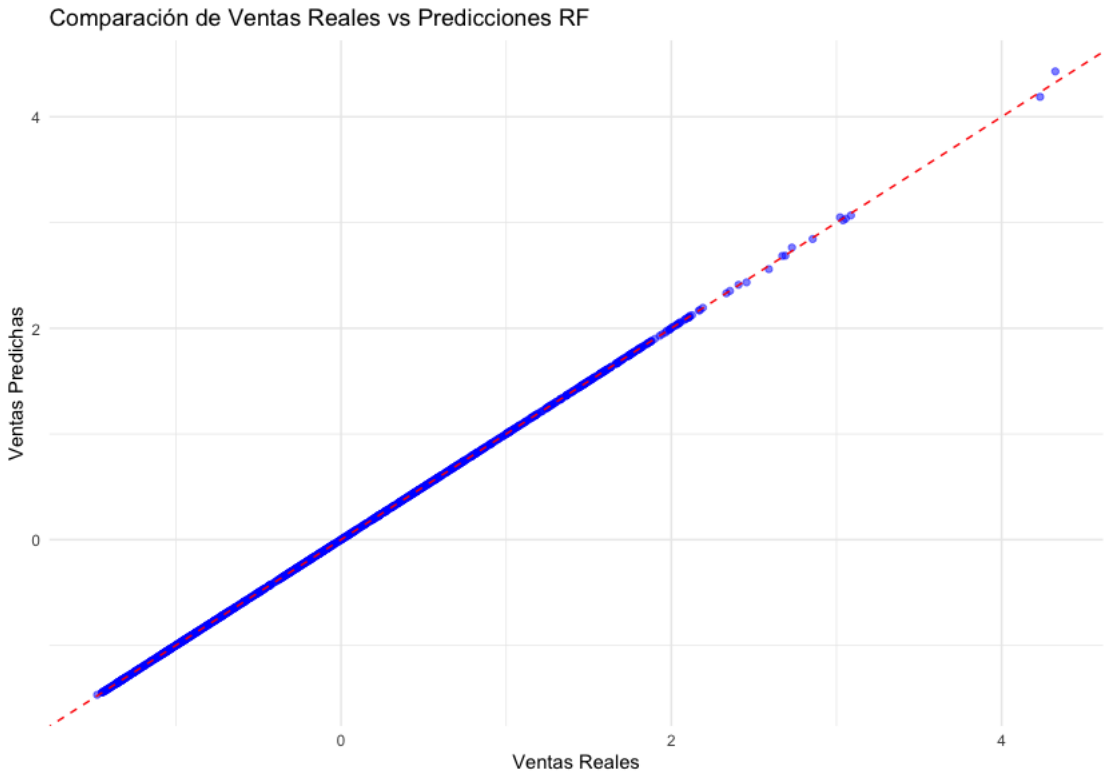
Una vez entrenado el modelo y decidido el número de variables a utilizar en las predicciones de Random Forest (51), procedemos a realizar las predicciones sobre el conjunto de prueba.

Figura 14. **Rendimiento de modelo Random Forest**

RMSE	Rsquared	MAE
0.0037963208	0.9999854865	0.0007751057

Al igual que sucedía en el conjunto de entrenamiento, las tres variables son extremadamente perfectas. De hecho, los resultados mejoran a los establecidos para "mtry" =51 que se mostraban en la figura anterior. Esto, lejos de significar ser un modelo perfecto, es una señal preocupante, ya que si los datos predichos se ajustan tan bien a los reales pueden indicar un sobreajuste y puede no generalizar correctamente a datos no conocidos previamente. Para verlo más claro representaremos gráficamente los valores reales y los predichos.

Figura 15. **Comparación de ventas reales y predicciones con Random Forest**



Como se puede apreciar en el gráfico, los datos reales y predichos se encuentran todos sobre la línea de regresión, lo que indica una precisión extraordinaria. Además, el modelo es consistentemente preciso a lo largo de todos los niveles de ventas ya que los puntos se disponen a lo largo de toda la línea de regresión. Sin embargo, como se ha mencionado arriba puede indicar un sobreajuste y una incapacidad del modelo de extrapolar los valores reales sobre valores desconocidos. Al existir 6435 registros, no se puede concluir que el sobreajuste sea evidente, por lo que conviene realizar alguna prueba al respecto. Para comprobar el sobreajuste, se pueden realizar dos cosas. La primera de ellas es la validación cruzada, pero ésta ya ha sido incorporada en el código antes de crear el modelo a través del método "repeatedcv" que consiste en la repetición del proceso de partición y entrenamiento para reducir la variabilidad en la estimación. La segunda opción es repetir el modelo utilizando otro método de partición. Si anteriormente utilizábamos el 80% de los datos para el conjunto de entrenamiento, conviene probar de nuevo con la configuración del 70% para ver si el modelo es consistente.

Figura 16. **Rendimiento del modelo de entrenamiento Random Forest (ajustado)**

mtry	RMSE	Rsquared	MAE
2	0.473695973	0.9240400	0.354250151
26	0.042040765	0.9979548	0.010388634
51	0.008596045	0.9998961	0.001288027

Cómo se puede observar, al nivel de las 51 variables, los valores de RMSE, Rsquared y MAE siguen siendo similares.
Lo que realmente conviene analizar son los resultados que obtiene el modelo sobre el conjunto de prueba.

Figura 17. **Rendimiento del Modelo Random Forest (ajustado)**

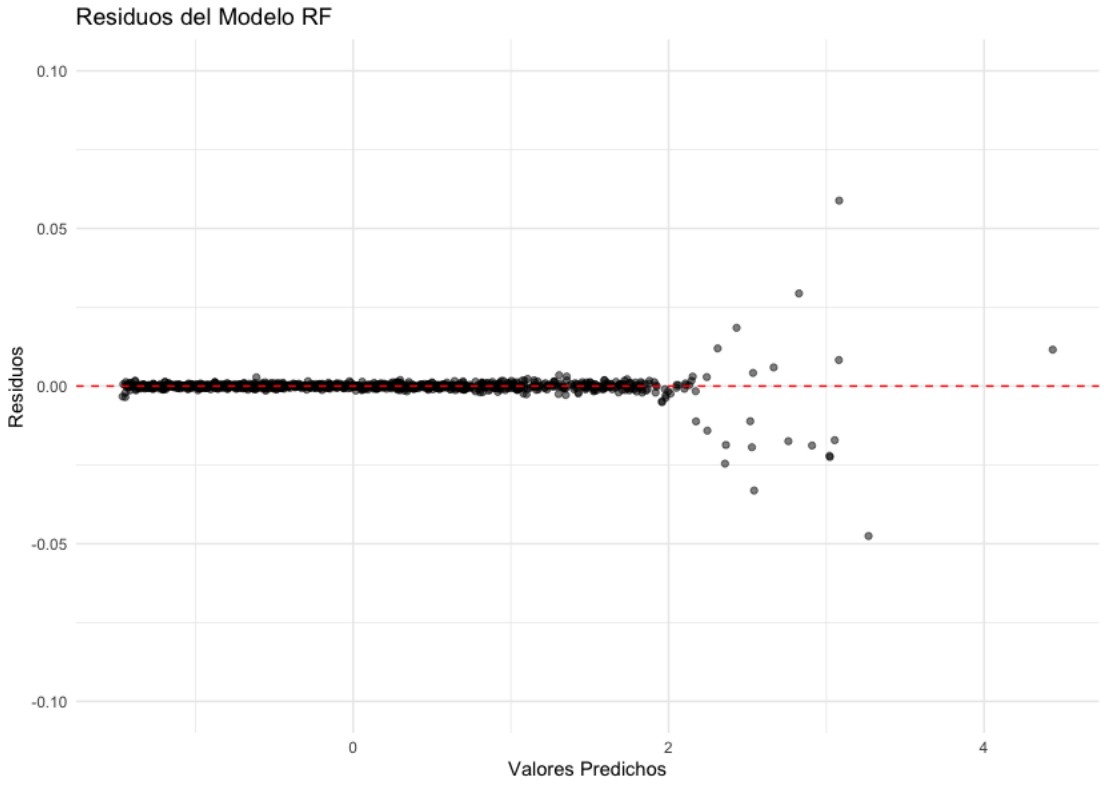
RMSE	Rsquared	MAE
0.0096205639	0.9999062296	0.0009545196

Como se puede apreciar, el modelo sigue mostrando la misma perfección en cuanto a la precisión de las predicciones y la capacidad de capturar todos los niveles de variabilidad. Tras haber realizado validación cruzada y haber probado los datos de entrenamiento con diferentes particiones sobre el conjunto de prueba, no se puede concluir que haya sobreajuste. En cualquier caso, tampoco se puede concluir que el modelo funcione a la perfección, ya que existen más

pruebas que se pueden realizar, pero a las que no tenemos acceso, y que ayudarían a comprobar si nos encontramos ante un modelo de predicción muy preciso o simplemente un modelo que tiene sobreajuste y no es capaz de extrapolar los datos a la realidad.

Por último, realizamos un análisis de residuos para evaluar la idoneidad del modelo.

Figura 18. **Análisis de Residuos de Modelo Random Forest**



Lo primero que hay que señalar es que los residuos parecen dispersarse a lo largo de la línea de cero, lo que es una buena señal. Además, el gráfico indica que no existe heterocedasticidad. Esto es, la variación desigual de los residuos a lo largo del rango de predicciones. Se encuentran unos pocos valores atípicos que pueden deberse a diversos factores y que sin duda serán una ventaja más a la hora de analizar los outputs que Walmart obtenga de la implementación del modelo. Estos valores atípicos representan un área en el que se puede profundizar para mejorar el modelo.

3.3. RESULTADOS

Tras haber estudiado el rendimiento y la capacidad predictiva de tres modelos distintos, se pueden sacar las siguientes conclusiones. En primer lugar, el modelo de regresión lineal es el que peor capacidad predictiva tiene. Ya sea por la complejidad de los datos que se le han imputado o por el problema de sobreajuste, los datos predichos no se acercan mucho a los verdaderos. Estos se dispersan a lo largo de todo el grafico alejándose de la línea de regresión y la línea de identidad. Tiene unos valores muy altos en cuanto a errores y muy bajos en cuanto en cuanto a capacidad explicativa. Los modelos que mejor funcionan son KNN y Random Forest. KNN, por su parte, consigue explicar el 96% de la variabilidad del modelo y muestra un bajo valor de errores, lo que implica que los valores reales y los predichos se asemejan y, por tanto, el modelo es muy preciso. Sin embargo, sigue habiendo ciertos valores que se alejan de la línea de identidad, sobre todo a medida que se avanza en los distintos niveles de venta. Por tanto, para las ventas de valor bajo es capaz de predecir de forma precisa, pero cuando van aumentando los niveles el valor de las predicciones se va alejando de los valores reales. Para una compañía como Walmart, en la que los niveles de venta suelen ser muy elevados y necesita saber el valor de ventas en todos los valores, esto es un problema.

Por último, el modelo de Random Forest es el que más precisión tiene. Todos los valores se concentran a lo largo de la línea de identidad y mantiene la precisión para todos los niveles de venta. Cómo se ha dicho, puede deberse a un sobreajuste, pero si la empresa prueba el modelo con otro conjunto de datos podrá comprobar si de verdad se trata de un problema de este tipo o confirmar que el modelo tiene una capacidad predictiva casi perfecta. Los errores son muy bajos y el modelo explica casi el 100% de la variabilidad de los datos.

Si bien es cierto que a nivel de coste computacional el modelo más barato sería el de regresión lineal, Walmart es una empresa que desde sus orígenes ha apostado por las soluciones tecnológicas de análisis de datos. Es por ello por lo que, además de contar con inmensos volúmenes de datos que reciben y analizan en tiempo real, la empresa es capaz de hacer frente a esos costes computacionales y apostar por un modelo que, pese a ser más caro, les reporte mejores resultados. Y ese modelo, como se ha desarrollado arriba, es el de Random Forest.

4. ALGORITMOS DE MEJORA DEL MODELO

Los algoritmos de regresión lineal, KNN y Random Forest son técnicas de aprendizaje automático que pueden considerarse un subcampo dentro de la Inteligencia Artificial. Al fin y al cabo, la Inteligencia Artificial son un conjunto de algoritmos que permite emular la inteligencia humana en algunos aspectos. Sin embargo, sí que es necesario recalcar que estas técnicas no son las más avanzadas que pueden producirse mediante Inteligencia Artificial. Existen otras que conllevan mayor entrenamiento algorítmico y coste computacional, pero que combinan la base de los modelos expuestos con una mayor precisión y acierto.

A continuación, se exponen algunas de estas técnicas avanzadas que optimizarían el modelo que mejor rendimiento ha aportado, el de Random Forest.

4.1. Deep Learning

Las redes neuronales profundas pueden ser cruciales para la obtención de características complejas. Para nuestro caso, el dataset de Walmart, únicamente contábamos con unas pocas variables dado que se trata de una simulación. Sin embargo, en un caso real, la empresa contará con muchas más métricas y características a las que debe dar el peso que tienen. Utilizando algoritmos de Deep learning como redes neuronales se pueden extraer esas características y su importancia y aporte al modelo, de forma que posteriormente se incluya como información de entrada al modelo de Random Forest.

4.2. Optimización Hiperparamétrica Automatizada

Los hiperparámetros son aquellos parámetros que se configuran de manera previa al entrenamiento del modelo de manera manual. En el modelo de Random Forest, los hiperparamétros serían el número de árboles en el bosque o el número de características utilizadas para la división.

Automatizarlos es útil porque se realiza de forma automática, valga la redundancia. Esto quiere decir que no es el analista el que tiene que ir probando con distintos números, si no que el propio algoritmo se va ajustando hasta encontrar la combinación más eficiente. Sin embargo, hay que tener en cuenta que existe el riesgo de sobreajuste sobre el modelo de validación, ya que la optimización podría acabar siendo perfecta en este campo en lugar de mejorar la optimización.

Algunas técnicas de optimización hiperparamétrica automatizada son Grid Search y Random Search.

4.2.1. Grid Search

Grid Search es una técnica que prueba continuamente una serie de valores de hiperparámetros especificados previamente y evalúa el rendimiento del modelo para cada combinación posible. Lo implementamos sobre nuestra base de datos y el código de random forest que teníamos para comprobar si optimiza el modelo.

Figura 19. **Modelo Random Forest optimizado con Grid Search**

```
      RMSE   Rsquared        MAE
0.2938244 0.9478039 0.2063490
```

Como se puede apreciar, en este caso Grid Search no ha mejorado los resultados del modelo anterior de Random Forest. Esto se puede deber a diferentes causas. Una de ellas es que el modelo anterior fuera óptimo que complicarlo con nuevas técnicas haya hecho perder eficacia en las predicciones. Otra de las razones es el tipo de Grid Search utilizado, en el que no se han seleccionado los valores óptimos. Por ello es razonable utilizar Random Search, ya que, a diferencia de Grid Search en el que había que especificar previamente el rango de valores para los hiperparámetros, utiliza combinaciones aleatorias.

4.2.2. Random Search

Tal y como se explicaba al final del apartado anterior, Random Search es razonable utilizarlo cuando el rango de hiperparámetros es desconocido, al igual que su impacto sobre el modelo de Random Forest. En este caso, lo ponemos en práctica para comprobar si mejora la capacidad predictiva del modelo de Random Forest, el que hasta ahora era el más preciso de todos los creados.

Figura 20. **Rendimiento del conjunto de entrenamiento de Random Forest (con Random Search)**

mtry	RMSE	Rsquared	MAE
3	0.315101892	0.9443293	0.216227751
4	0.240921122	0.9553005	0.148364804
6	0.177964316	0.9699420	0.092310862
9	0.133497185	0.9820452	0.058327147
10	0.122462550	0.9847488	0.051011971
14	0.089663948	0.9915188	0.030868201
18	0.067759680	0.9949288	0.019639185
22	0.051523987	0.9968760	0.012994417
27	0.037383103	0.9982791	0.008156342
29	0.032730476	0.9986347	0.006781827
32	0.026783457	0.9990455	0.005169445
37	0.019159819	0.9995022	0.003311426
41	0.014899023	0.9997005	0.002358577
44	0.012279854	0.9998011	0.001807998
45	0.011570274	0.9998229	0.001650893
50	0.008240643	0.9999071	0.001143086
51	0.007839842	0.9999140	0.001108368

Tras entrenar el modelo utilizando Random Search, observamos que se han realizado numerosas iteraciones probando cada número de variables ("mtry") probando las combinaciones más eficientes para el conjunto de entrenamiento. Siguiendo estos datos, el número óptimo de variables o características con las que debe contar el modelo es 51. El algoritmo de Random Search, además, ha ajustado el valor de los hiperparámetros que se utilizarán sobre el conjunto de prueba.

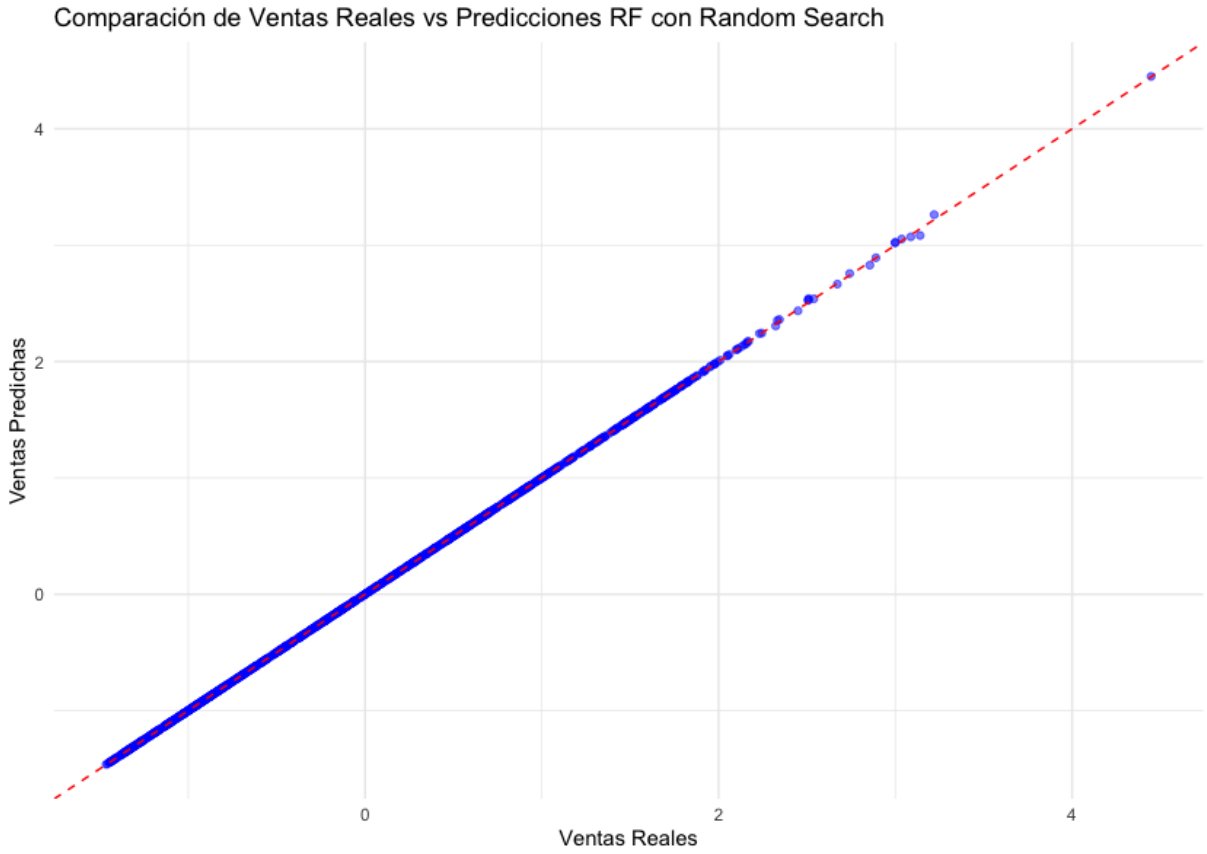
Figura 21. **Rendimiento del modelo Random Forest (con Random Search)**

RMSE	Rsquared	MAE
0.0029903127	0.9999909966	0.0007542324

Estos resultados hay que ponerlos en comparación con las métricas obtenidas en el modelo de Random Forest normal, expuestos en la Figura 13. Como se puede apreciar, el RMSE es menor, Rsquared mayor y MAE menor que antes de utilizar Random Search. Todo ello quiere decir que el modelo es aún más preciso que antes y, aunque aún existe el riesgo de sobreajuste, el modelo es casi perfecto.

Si representamos gráficamente la comparativa de resultados entre los valores reales y los predichos quedaría como se muestra en el siguiente gráfico.

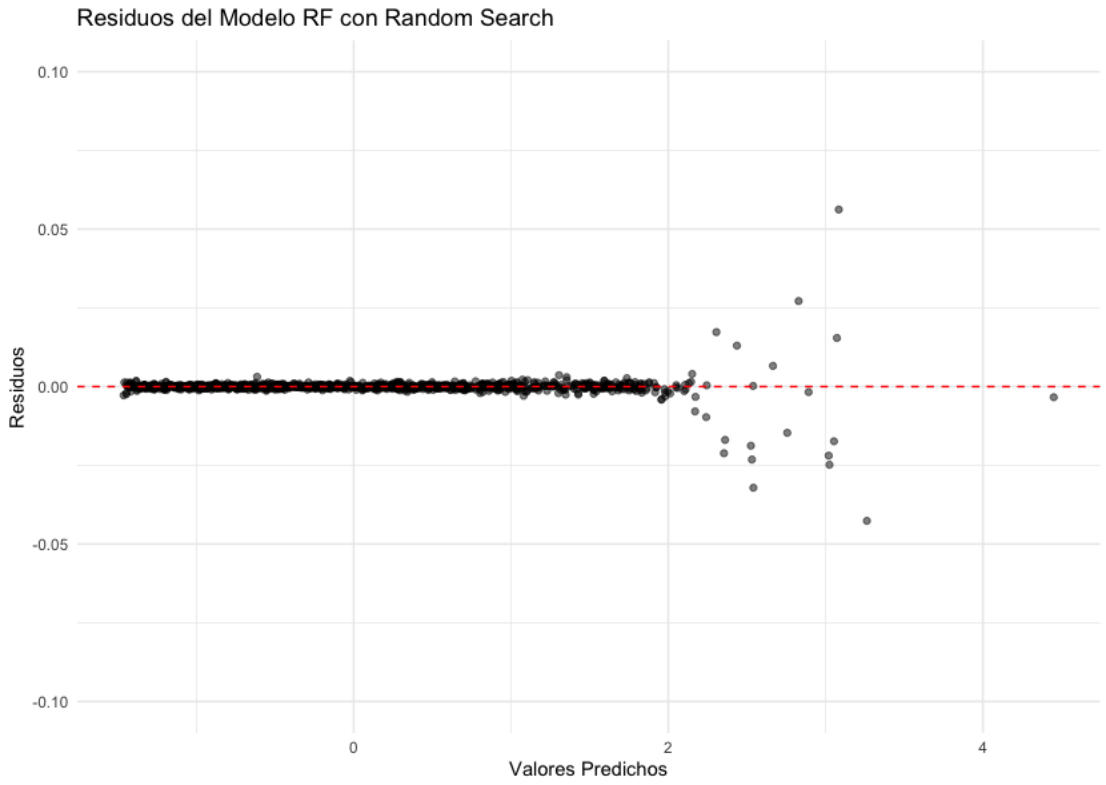
Figura 22. **Ventas reales y ventas predichas del modelo Random Forest (con Random Search)**



Como se puede observar en el gráfico, todos los valores se encuentran sobre la línea de identidad y en completa armonía. Como sucedía antes, a medida que subes los niveles de ventas los resultados se dispersan un poco, pero siguen concentrándose en torno a la línea.

Para terminar de analizar el algoritmo, se realiza un análisis de residuos.

Figura 23. **Residuos del Modelo Random Forest (con Random Search)**



Al analizar el gráfico de residuos y comparándolo con el gráfico de la Figura 18 en el que se mostraban los residuos del modelo de Random Forest en el que no se había utilizado ningún algoritmo de optimización hiperparamétrica, lo que se observa es que en este caso los outliers resultan ser ligeramente menos extremos y concentra más los residuos en torno a la línea de cero que el otro modelo. Sin embargo, no es una mejora muy significativa ya que los outliers siguen existiendo y como mucho rebajan unas milésimas. Esto ya se podía ver comparando los resultados de rendimiento de ambos modelos, en los que las métricas de RMSE, Rsquared y MAE eran muy similares. En cualquier caso, Random Search aumenta la precisión del modelo y mejora el rendimiento.

Como conclusión, si bien el modelo de Random Forest era casi perfecto y contaba con una gran precisión, añadirle el algoritmo de Random Search le proporciona un ligero mayor rendimiento que se observa tanto en la identidad entre valores reales y predichos como en el análisis de residuos.

Es, por tanto, una mejora del modelo que sin duda le añade más valor. Sin embargo, durante la ejecución del algoritmo he podido observar el elevado coste computacional y de tiempo que

conlleva, por lo que se deberá analizar si merece la pena utilizarlo en modelos que ya arrojan resultados casi perfectos, dado que la mejora será mínima.

4.3. Interpretación de Modelos con IA Explicable

La Interpretación de modelos con IA Explicable o XAI por sus siglas inglés es un cojunto de métodos de Inteligencia Artificial dirigidos a hacer más interpretable por los humanos los modelos construidos con algoritmos de Inteligencia Artificial. Proporcionan explicaciones claras y comprensibles sobre cómo toman las decisiones. Implica varias herramientas que ayudan a entender mejor cuales son las métricas de decisión. Un resumen de las técnicas de XAI se muestran en la siguiente imagen.

De acuerdo con Adadi y Berrada (2019) , existen principalmente dos tipos de tecnicas: model-specific y model-agnostic. Las del primer tipo son aquellas que están ligadas a un único tipo de modelo algorítmico, mientras que las del segundo pueden ser utilizadas sobre cualquier tipo de modelo.
Para el modelo de Random Forest, se va a utilizar el estudio de PDP de las características del modelo para conocer más a fondo su utilidad en la predicción.

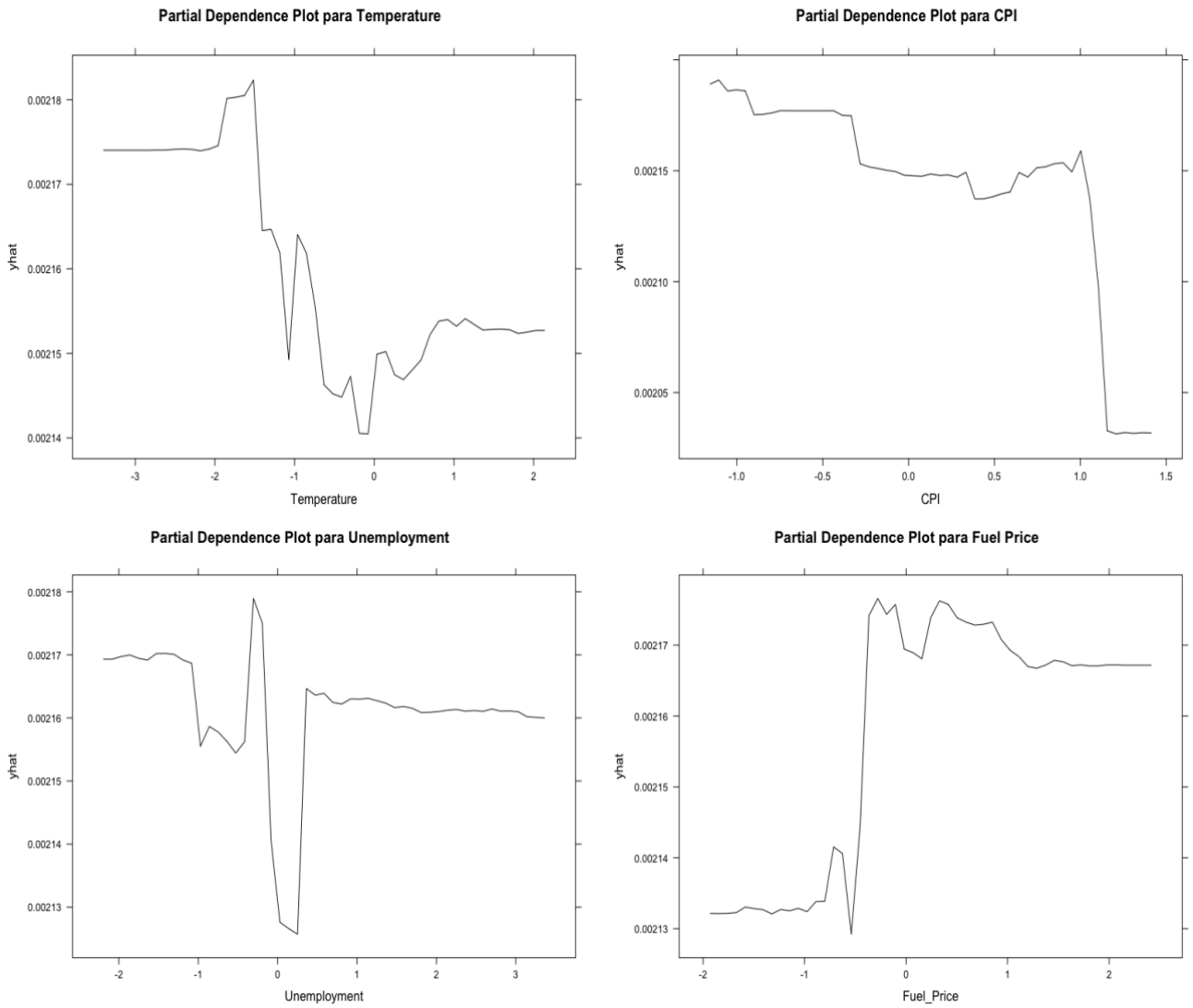
4.3.1. Partial Dependance Plot

Tal y como se ha descrito en la Figura 24, la técnica Partial Dependance Plot (PDP) proporciona una visión de forma gráfica de la contribución que tienen variables independientes al modelo. Ilustra cómo se espera que varíe la predicción promedio del modelo al cambiar los valores de las características de interés, manteniendo constantes todas las demás características. En modelos complejos, esta técnica puede ser muy útil para ir analizando la forma en la que las variables se relacionan, y como varían a medida que el modelo se vuelve más complejo.

El PDP ayuda en la interpretación de los modelos complejos, en los que hay variables independientes que tienen un peso específico sobre el modelo. Sin embargo, esta técnica tiene limitación, ya que asume la independencia entre las características y puede ser engañoso si existe una fuerte correlación entre variables.

A continuación, se expone la representación gráfica de PDP de las variables independientes del modelo.

Figura 24. **Partial Dependance Plot de las características**



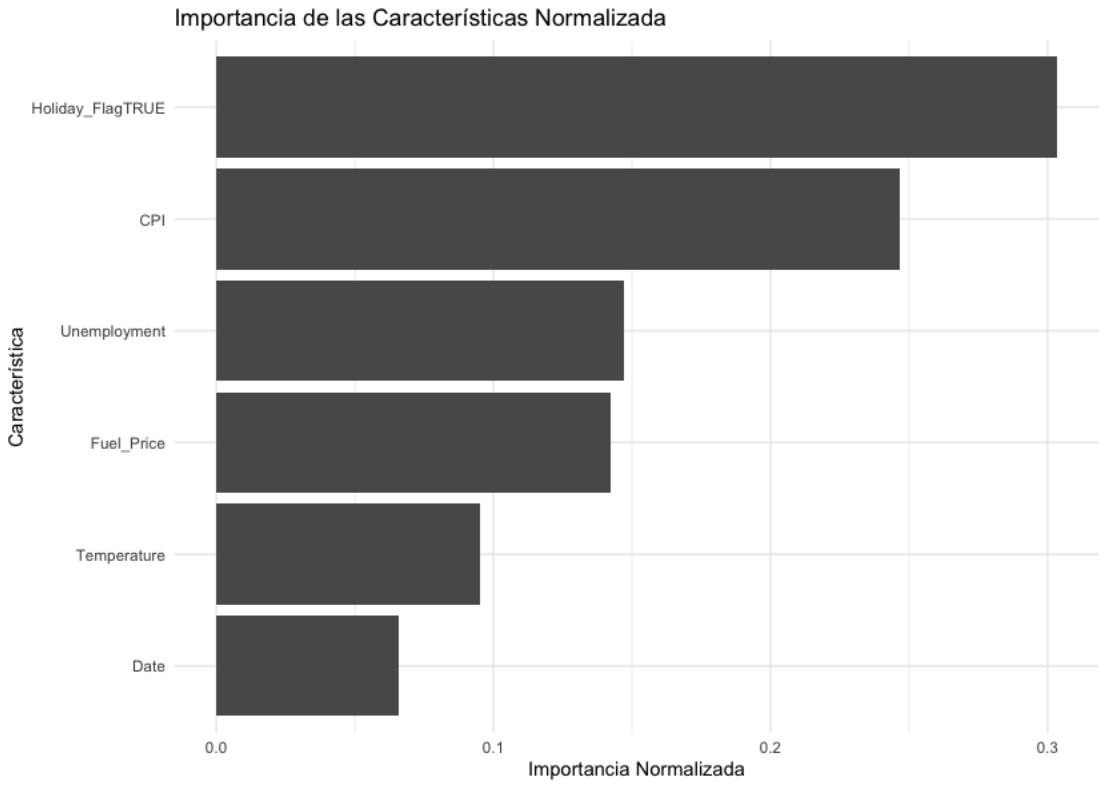
Analizando brevement cada uno de los gráficos, podemos extraer las siguientes conclusiones. En lo referente a la Temperature, hay cierta variabilidad en la influencia de esta variable sobre el modelo. Se produce un pico en el que parece que hay un cambio en la tendencia predictora de la variable.

En cuanto al CPI, las predicciones del modelo van disminuyendo a medida que el CPI aumenta, lo cual puede significar que a partir de un precio del indice de precios del consumidor, las ventas decaen.
Por su parte, Unemployment tiene una variabilidad especifica en dos picos, arriba y abajo. Existe una gran sensibilidad a los datos en una franja, que puede indicar que en determinados niveles de desempleo hay mas inestabilidad en cuanto a la precisión de la predicción.
En cuanto a Fuel Price, se observa otro pico ascendente para luego estabilizarse. La aportación de esta característica a la predicción aumenta a medida que aumenta el valor de la variable.

4.3.2. Importancia de las carácterísticas

La técnica de importancia sirve para determinar el peso que las variables tienen sobre el modelo de Random Forest, ya que, como se ha indicado arriba es model-specific para este tipo de modelos. En la siguiente figura se puede ver el desglose gráfico de estas características. Cabe mencionar que la variable Weekly_Sales, así como su versión estandarizada (sales_standar) han sido eliminadas del gráfico ya que, como es lógico, eran las variables más importantes dentro del dataset y convenía centrarse en las demás.

Figura 25. **Importancia de las características**



Como resultados, se aprecia que la variable más influyente es "Holiday_Flag", que recordemos que indicaba si el día en el que se registraba la venta era festivo. También es necesario puntualizar que los pocos días festivos (TRUE) que existen en la base de datos y su posible coincidencia con picos de venta altos, hace que el nivel de importancia incremente. El resto de variables siguen una lógica más lineal, en la que la temperatura y la fecha son poco relevantes y factores económicos como el CPI y la tasa de desempleo (Unemployment) influyen más en el volumen de ventas.

4.3.3. Resultados

Cómo se puede observar, la gran mayoría de los métodos expuestos funcionan explicando la proporción de valor que aportan distintas características dentro del modelo, de forma que se puede determinar las variables más influyentes y que la empresa debe considerar a la hora de afrontar un análisis y predicción de niveles de ventas, como es el caso de la técnica Importance.

Además, la utilización de PDP sirve para determinar cómo las predicciones podrían cambiar en promedio con los valores de cada característica.

En definitiva, son técnicas muy útiles que ayudan a entender mejor cómo el modelo determina sus resultados predictivos y las razones que existen detrás de ello, además de poder utilizarse para otros medios como, por ejemplo, afrontar temporadas de rebajas con resultados excelentes y comprendiendo los factores que movilizan el consumo.

5. CONCLUSIONES

Walmart es una empresa cuya gran ventaja competitiva es la cantidad de datos que son capaces de recopilar y analizar, extrayendo de éstos conclusiones que sirven para tomar decisiones estratégicas o arreglar posibles errores que se encuentren en la cadena de suministros o de venta. Tan importante es la capacidad de extracción como la correcta utilización de las herramientas de Big Data para darle buen uso a dicha información. En este punto, y para el tema que nos ocupa en este trabajo de fin de grado, es importante conocer bien a qué modelos de Machine Learning se puede acudir para optimizar el proceso de venta.

Tras haber experimentado con distintos modelos y, como es evidente, el más complejo es el que mejor resultados da. El modelo de regresión lineal, al ser más simple, no era capaz de recopilar toda la información necesaria para extrapolar y generalizar las predicciones. Uno de los fallos que se pueden encontrar es que no se haya contado con una posible correlación no lineal entre variables que sí que muestran otros modelos. El modelo de KNN ajusta mucho mejor las predicciones a los valores reales, pero seguía mostrando fallos de precisión y una serie de outliers que no era capaz de asumir el modelo. Por su parte, el modelo de Random Forest resultó ser el mejor de todos, siendo capaz de explicar casi el 100% de la variabilidad y con el menor riesgo de error.

Sin embargo, son tantas las posibilidades y métodos que ofrece la Inteligencia Artificial, que quedarse con el simple modelo de Random Forest podría ser un error. La exploración de distintas técnicas puede llevar a la perfección del modelo. En este caso, aplicando el método de optimización hiperparamétrica de Random Search obtuvimos una mejora en todas las métricas de rendimiento. Es importante, en cualquier caso, tener en cuenta el coste computacional que conlleva cada una de las técnicas avanzadas y conocer muy bien el modelo del que se parte para considerar si de verdad es necesario llevar a cabo un proceso de optimización hiperparamétrica.

Si bien la construcción de modelos debe ser perfecta, el trabajo de interpretación de estos resultados parece incluso más importante. Los resultados que arrojan los modelos pueden ser complejos de entender y a simple vista pueden no decir nada, pero la aplicación de las distintas técnicas explicativas, así como técnicas de visualización pueden resultar cruciales para un

resultado óptimo. Lo positivo de estas técnicas explicativas es que permiten a las empresas de menor tamaño analizar los datos y sacar conclusiones sin la necesidad de contar con un analista de datos en su plantilla.

Tampoco se debe olvidar la posibilidad de que exista sobreajuste en los datos, y que el modelo entrenado arroje los mismos datos sobre el conjunto de prueba. Esto, en una empresa como Walmart con tanta cantidad de datos obtenidos por día, puede ser fácilmente comprobable utilizando otro conjunto de datos sobre el mismo modelo.

Como conclusión, las técnicas de IA y en concreto de Machine Learning, son esenciales en el proceso de venta y pueden resultar cruciales ante algún problema. La evolución de las técnicas permite a las empresas aprovechar grandes cantidades de datos y descubrir, por ejemplo, tendencias de compra. Ser capaz de responder en tiempo real ante fluctuaciones de mercado u ofrecer personalizaciones a los usuarios son solo algunas de las ventajas que ofrece la IA. Es un elemento clave en la búsqueda de ser el mejor del mercado.

Declaración de Uso de Herramientas de Inteligencia Artificial Generativa en Trabajos Fin de Grado

ADVERTENCIA: Desde la Universidad consideramos que ChatGPT u otras herramientas similares son herramientas muy útiles en la vida académica, aunque su uso queda siempre bajo la responsabilidad del alumno, puesto que las respuestas que proporciona pueden no ser veraces. En este sentido, NO está permitido su uso en la elaboración del Trabajo fin de Grado para generar código porque estas herramientas no son fiables en esa tarea. Aunque el código funcione, no hay garantías de que metodológicamente sea correcto, y es altamente probable que no lo sea.

Por la presente, yo, Ignacio Peña Sanz, estudiante de Derecho y Business Analytics de la Universidad Pontificia Comillas al presentar mi Trabajo Fin de Grado titulado " Predicción de ventas utilizando métodos de Big data y su posterior tratamiento a través de Inteligencia Artificial", declaro que he utilizado la herramienta de Inteligencia Artificial Generativa ChatGPT u otras similares de IAG de código sólo en el contexto de las actividades descritas a continuación:

1. **Brainstorming de ideas de investigación:** Utilizado para idear y esbozar posibles áreas de investigación.
2. **Interpretador de código:** Para realizar análisis de datos preliminares.
3. **Corrector de estilo literario y de lenguaje:** Para mejorar la calidad lingüística y estilística del texto.
4. **Sintetizador y divulgador de libros complicados:** Para resumir y comprender literatura compleja.
5. **Revisor:** Para recibir sugerencias sobre cómo mejorar y perfeccionar el trabajo con diferentes niveles de exigencia.

Afirmo que toda la información y contenido presentados en este trabajo son producto de mi investigación y esfuerzo individual, excepto donde se ha indicado lo contrario y se han dado los créditos correspondientes (he incluido las referencias adecuadas en el TFG y he explicitado para que se ha usado ChatGPT u otras herramientas similares). Soy consciente de las implicaciones académicas y éticas de presentar un trabajo no original y acepto las consecuencias de cualquier violación a esta declaración.

Fecha: 23 de abril de 2024

Firma: Ignacio Peña Sanz

Bibliografía

Adadi, A., & Berrada, M. (2019). *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI).* Obtenido de ResearchGate: https://www.researchgate.net/publication/327709435_Peeking_Inside_the_Black-Box_A_Survey_on_Explainable_Artificial_Intelligence_XAI

AEMR (Asociación Española de Marketing Relacional). (2002). *III Estudio de CRM en España.* Obtenido de https://www.fecemd.org/ar-chivos/aemr-estudio2002.pdf

BBC. (21 de Marzo de 2018). *5 claves para entender el escándalo de Cambridge Analytica que hizo que Facebook perdiera US$37.000 millones en un día*. Obtenido de BBC: https://www.bbc.com/mundo/noticias-43472797

BBVA. (2021). *Las cinco uves del Big Data.* Obtenido de https://www.bbva.com/es/innovacion/las-cinco-uves-del-big-data/

Bron Fonseca, B., & Mar Cornelio, O. (Enero - Abril de 2022). Sistemas de Recomendación para la toma de decisiones. Estado del Arte. *UNESUM - Ciencias: Revista Científica Multidisciplinaria, 6*(1), 149-164.

Cao, P. (2021). Big Data in Customer Acquisition and Retention for eCommerce - Taking Walmart as an Example. *Proceedings of the 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021).* Atlantis Press International B.V.

Dyer, J., & Gregersen, H. (29 de Mayo de 2018). *ServiceNow, Workday and Salesforce are driving Digital Transformation*. Obtenido de Forbes.com: https://www.forbes.com/sites/innovatorsdna/2018/05/29/servicenow-workday-and-salesforce-are-driving-digital-transformation/?sh=691282ee313b

European Commission. (2022). *AI Watch: Defining Artificial Intelligence. Towards an operational definition and taxonomy of artificial intelligence.* Obtenido de European Union Publications Office: https://data.europa.eu/doi/10.2760/382730

Ferreiro Piñeiro, I., & Alonso Varona, L. (enero-junio de 2019). Big Data y Gestión Empresarial. *Revista técnica de la empresa de telecomunicaciones de Cuba, 15*(1), 27-32.

García, F., & Gil, A. (2020). *Personalización de Sistemas de Recomendación.* Universidad de Salamanca, Dpto. Informática y Automática.

Gates, B. (2 de abril de 2023). La edad de la Inteligencia Artificial ha comenzado. *La Vanguardia*, pág. 32.

Humberto, B.-S., Ramon, C.-M., & Emili, G.-T. (2015). Business Model Evaluation: Quantifying Walmart's Sources of Advantage. *Strategic Entrepreneurship Journal, 9*(1), 12-33.

Jiménez Estrada, V., & Gómez Herrera, G. (2021). *Propuesta de desarrollo de un sistema de predicción de ventas usando algoritmos de Inteligencia Artificial.* Universidad Estatal de Milagro, Facultad de ciencias e ingeniería, Milagro.

Joshi, A., Spilbergs, A., & Mikelsone, E. (15 de August de 2023). *Impact of Digitalization, Big Data Analytics and Artificial Intelligence on Business Transformation in IT Companies: Literature Review and Conceptual Framework.* Obtenido de EasyChair Preprints: https://easychair.org/publications/preprint_download/KJwB

Kotler, P., & Keller, K. L. (2015). *Marketing Management.* Pearson.

Li, Y. (26 de Noviembre de 2018). *Deep Reinforcement Learning: An overview.* Obtenido de Arxiv: https://doi.org/10.48550/arXiv.1810.06339

López, F. G. (2017). *Características de Big Data.* Obtenido de Observatorio de BI & Analytics: https://spaceanalytics.blogspot.com/2016/05/caracteristicas-de-big-data.html

Mahmoud, & Rasha. (Diciembre de 2019). *Discount Store*. Obtenido de Retail Dogma: https://www.retaildogma.com/discount-store/

Marr, B. (January de 2017). *Really Big Data At Walmart: Real-Time Insights From Their 40+ Petabyte Data Cloud*. Obtenido de Forbes.com: https://www.forbes.com/sites/bernardmarr/2017/01/23/really-big-data-at-walmart-real-time-insights-from-their-40-petabyte-data-cloud/?sh=24e576916c10

McCarthy, J. (10 de 09 de 2007). *From here to human-level AI.* Obtenido de Science Direct: https://pdf.sciencedirectassets.com/271585/1-s2.0-S0004370207X03151/1-s2.0-S0004370207001476/main.pdf?X-Amz-Security-Token=IQoJb3JpZ2luX2VjEHoaCXVzLWVhc3QtMSJHMEUCIQDUzZ2YQDGvQPm1ZP5V0wTkYS95Ls9p4WqEWqVgSgJB7AIgFZEXL4%2BG5O0FTkHHRgeVpEfnSQZ4o%2BqfU1DVXEou

Moreno Espinosa, P., Abdulsalam Alsarayreh, R., & Figuereo Benitez, J. (Enero - Julio de 2024). El Big Data y la inteligencia artificial como soluciones a la desinformación. *Revista Doxa Comunicación*(38), 437 - 453.

Moreno, J. P. (2014). Una aproximación a Big Data. *Revista de Derecho UNED*(14), 471 - 505.

OECD. (Mayo de 2019). *OECD Economic Outlook.* Obtenido de https://www.oecd-ilibrary.org//sites/b2e897b0-en/1/2/2/index.html?itemId=/content/publication/b2e897b0-en&_csp_=d2743ede274dd564946a04fc1f43d5dc&itemIGO=oecd&itemContentType=book#section-d1e3167

Pastor, J. (23 de Noviembre de 2017). *Da igual que desactives el GPS: Google sigue recolectando tu ubicación en Android*. Obtenido de Xataka:

https://www.xataka.com/privacidad/da-igual-que-desactives-el-gps-google-sigue-recolectando-tu-ubicacion-en-android

Peco, R. (4 de Mayo de 2018). No es una leyenda urbana, Google escucha a través de los teléfonos. *La Vanguardia*.

Prometeus Global Solutions. (19 de Febrero de 2019). *Volumen, Variedad, Velocidad, Veracidad y Valor: las 5 dimensiones del Big Data.* Obtenido de https://prometeusgs.com/volumen-variedad-velocidad-veracidad-y-valor-las-5-dimensiones-del-big-data-la/

Sandoval, L. J. (Diciembre de 2018). Algoritmos de aprendizaje automático para análisis y predicción de datos. *Revista Tecnológica*(11), 36 - 40.

Schroeck, M., Shockley, R., Smart, D., Romero-Morales, P., & Tufano, P. (2014). *Analytics: el uso de big data en el mundo real: cómo las empresas más innovadoras extraen valor de datos inciertos.* IBM, Institute for Business Value. Saïd Business School.

The Economist. (6 de Mayo de 2017). The world´s most valuable resource is no longer oil, but data. *The Economist*.

The Editors of Encyclopaedia. (29 de September de 2023). *Walmart. American company*. Obtenido de Britannica: https://www.britannica.com/topic/Walmart

Toonders, Y. J. (2014). *Data is the New Oil of the Digital Technology.* Obtenido de Wired.com: https://www.wired.com/insights/2014/07/data-new-oil-digital-economy/

Anexo. Código de R

Carga de librerías

```
library(ggplot2)
library(caret)
library(ROCR)
library(GGally)
library(rpart)
library(rpart.plot)
library(caretEnsemble)
library(tictoc)
library(MASS)
library(partykit)
library(grid)
library(randomForest)
library (lubridate)
library (ggcorrplot)
library (dplyr)
library (broom)
library(pdp)
library(shap)
```

Preprocesamiento de datos y análisis de variables

```
walmart <- read.csv("Walmart.csv")
View (walmart)

#REVISAMOS LAS VARIABLES
summary (walmart)
str(walmart)
walmart$Holiday_Flag<- as.logical(walmart$Holiday_Flag)
walmart$Date <- as.Date(walmart$Date, format = "%d-%m-%Y")
walmart$Store <- as.factor(walmart$Store)
walmart$sales_standar <- scale(walmart$Weekly_Sales)

# Normalización de las variables
preProcess_range_model <- preProcess(walmart[, c("Weekly_Sales", "Temperature", "Fuel_Price", "CPI", "Unemployment")], method=c("center", "scale"))
```

```
walmart_normalized <- predict(preProcess_range_model, walmart)

# Matriz de correlación entre variables numéricas
cor_matrix <- cor(walmart_normalized[, c("Weekly_Sales", "Temperature", "Fuel_Price", "CPI", "Unemployment", "Holiday_Flag")])
ggcorrplot(cor_matrix, lab = TRUE)
```

Modelo de regresión lineal

```
set.seed(123)
trainIndex <- createDataPartition(walmart_normalized$Weekly_Sales, p = .8, list = FALSE)
train_set <- walmart_normalized[trainIndex, ]
test_set <- walmart_normalized[-trainIndex, ]
train_control_glm <- trainControl(
        method = "repeatedcv",
        number = 10,
        repeats = 3)

modelo <- lm(Weekly_Sales ~ Temperature + Fuel_Price + CPI + Unemployment + Holiday_Flag, data = train_set)
summary(modelo)

# Entrenamiento del modelo
modelo1 <- train(
        Weekly_Sales ~ CPI + Unemployment + Holiday_Flag,
        data = walmart_normalized,
        method = "glm",
        preProcess = c("center", "scale"),
        trControl = train_control_glm,
        metric = "RMSE")

# Resumen del modelo entrenado
summary(modelo1)
print(modelo1)

# Predicciones en el conjunto de prueba
predictions_glm <- predict(modelo1, newdata = test_set)
```

```
# Comparación de las predicciones con los valores reales
performance_glm <- postResample(pred = predictions_glm, obs = test_set$Weekly_Sales)
performance_glm

results_glm <- data.frame(Real = test_set$Weekly_Sales, Predicted = predictions_glm)

# Gráfico de dispersión
ggplot(results_glm, aes(x = Real, y = Predicted)) +
        geom_point(alpha = 0.7) +
        geom_smooth(method = "lm", se = FALSE, color = "red") +
        geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "blue") +
        labs(title = "Comparación de Ventas Reales vs Predichas",
        x = "Ventas Reales (Normalizadas)",
        y = "Ventas Predichas (Normalizadas)") +
        theme_minimal()
```

Modelo de KNN

```
repeats = 3
numbers = 10
tunel = 5

RNGkind("Super", "Inversion", "Rounding")
set.seed(100)

x = trainControl(method = "repeatedcv",
            number = numbers,
            repeats = repeats)

knntrain <- train(Weekly_Sales~. , data = train_set,
             method = "knn",
             preProcess = c("center","scale"),
             trControl = x,
             metric = "RMSE",
             tuneLength = tunel)

knntrain
plot(knntrain)
knn_testclass <-predict(knntrain, newdata = test_set)
```

```
knn_performance <- postResample(pred = knn_testclass, obs = test_set$Weekly_Sales)
print(knn_performance)

results_knn <- data.frame(Real = test_set$Weekly_Sales, Predicted = knn_testclass)

# Gráfico de dispersión
ggplot(results_knn, aes(x = Real, y = Predicted)) +
  geom_point(color = "blue", alpha = 0.5) +
  geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
  labs(title = "Comparación de Ventas Reales vs Predicciones KNN",
       x = "Ventas Reales",
       y = "Ventas Predichas") +
  theme_minimal()

Modelo Random Forest
set.seed(123)
train_control_rf <- trainControl(method = "repeatedcv", number = 10, repeats = 3)

# Entrenamiento del modelo de Random Forests
rf_model <- train(
            Weekly_Sales ~ .,
            data = train_set,
            method = "rf",
            trControl = train_control_rf,
            metric = "RMSE",
            ntree = 100)
print(rf_model)

actual_values_rf <- test_set$Weekly_Sales
postResample(pred = predictions_rf, obs = actual_values_rf)

# Predicciones en el conjunto de prueba
predictions_rf <- predict(rf_model, newdata = test_set)
results_rf <- data.frame(Real = test_set$Weekly_Sales, Predicted = predictions_rf)
```

```
# Gráfico de dispersión
ggplot(results_rf, aes(x = Real, y = Predicted)) +
        geom_point(color = "blue", alpha = 0.5) +
        geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
        labs(title = "Comparación de Ventas Reales vs Predicciones RF",
        x = "Ventas Reales",
        y = "Ventas Predichas") +
        theme_minimal()

results_rf$Residuals <- results_rf$Real - results_rf$Predicted
ggplot(results_rf, aes(x = Predicted, y = Residuals)) +
        geom_point(alpha = 0.5) +
        geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
        labs(title = "Residuos del Modelo RF",
        x = "Valores Predichos",
        y = "Residuos") +
        ylim(c(-0.1, 0.1)) +
        theme_minimal()

#Otro conjunto de entrenamiento
set.seed(123)
index_seg <- createDataPartition(walmart$Weekly_Sales, p = 0.7, list = FALSE)
train_set_2 <- walmart_normalized[index_seg, ]
test_set_2 <- walmart_normalized[-index_seg, ]
rf_model_2 <- train(
        Weekly_Sales ~ .,
        data = train_set_2,
        method = "rf",
         trControl = train_control_rf,
        metric = "RMSE",
        ntree = 100)

print(rf_model_2)
predictions_rf_2 <- predict(rf_model_2, newdata = test_set_2)
actual_values_rf_2 <- test_set_2$Weekly_Sales
postResample(pred = predictions_rf_2, obs = actual_values_rf_2)
```

Grid Search

```
set.seed(123)
train_control_rf_gs <- trainControl(method = "repeatedcv", number = 10, repeats = 3, search = "grid")

tune_grid <- expand.grid(mtry = c(2, round(sqrt(ncol(train_set))), round(ncol(train_set)/3)))

rf_model_gs <- train(
        Weekly_Sales ~ .,
        data = train_set,
        method = "rf",
         trControl = train_control_rf_gs,
        metric = "RMSE",
        tuneGrid = tune_grid)

print(rf_model_gs)

# Predicciones en el conjunto de prueba
predictions_rf_gs <- predict(rf_model_gs, newdata = test_set)
actual_values_rf_gs <- test_set$Weekly_Sales
postResample(pred = predictions_rf_gs, obs = actual_values_rf_gs)

results_rf_gs <- data.frame(Real = test_set$Weekly_Sales, Predicted = predictions_rf_gs)

# Gráfico de dispersión
ggplot(results_rf_gs, aes(x = Real, y = Predicted)) +
        geom_point(color = "blue", alpha = 0.5) +
        geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
        labs(title = "Comparación de Ventas Reales vs Predicciones RF",
        x = "Ventas Reales",
        y = "Ventas Predichas") +
        theme_minimal()

# Residuos
results_rf_gs$Residuals <- with(results_rf_gs, Real - Predicted)

results_rf_gs$Residuals <- results_rf_gs$Real - results_rf_gs$Predicted
```

```
ggplot(results_rf_gs, aes(x = Predicted, y = Residuals)) +
        geom_point(alpha = 0.5) +
        geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
        labs(title = "Residuos del Modelo RF",
        x = "Valores Predichos",
        y = "Residuos") +
        theme_minimal()
```

Random Search

```
set.seed(123)

train_control_rf_rs <- trainControl(method = "repeatedcv",
                            number = 10,
                            repeats = 3,
                            search = "random")

rf_model_rs <- train(
        Weekly_Sales ~ .,
        data = train_set,
        method = "rf",
        trControl = train_control_rf_rs,
        metric = "RMSE",
        tuneLength = 20)

print(rf_model_rs)

# Predicciones en el conjunto de prueba
predictions_rf_rs <- predict(rf_model_rs, newdata = test_set)
actual_values_rf_rs <- test_set$Weekly_Sales
postResample(pred = predictions_rf_rs, obs = actual_values_rf_rs)

results_rf_rs <- data.frame(Real = test_set$Weekly_Sales, Predicted = predictions_rf_rs)
```

```
# Gráfico de dispersión
ggplot(results_rf_rs, aes(x = Real, y = Predicted)) +
        geom_point(color = "blue", alpha = 0.5) +
        geom_abline(intercept = 0, slope = 1, linetype = "dashed", color = "red") +
        labs(title = "Comparación de Ventas Reales vs Predicciones RF con Random Search",
        x = "Ventas Reales",
        y = "Ventas Predichas") +
        theme_minimal()

# Gráfico de residuos
results_rf_rs$Residuals <- results_rf_rs$Real - results_rf_rs$Predicted
ggplot(results_rf_rs, aes(x = Predicted, y = Residuals)) +
        geom_point(alpha = 0.5) +
        geom_hline(yintercept = 0, linetype = "dashed", color = "red") +
        labs(title = "Residuos del Modelo RF con Random Search",
        x = "Valores Predichos",
        y = "Residuos") +
        ylim(c(-0.1, 0.1)) +
        theme_minimal()
```

Técnicas de XAI

PDP

```
partial_rf_CPI <- partial(rf_model, pred.var = "CPI")
plotPartial(partial_rf_CPI, main = "Partial Dependence Plot para CPI")

partial_rf_Unemploy <- partial(rf_model, pred.var = "Unemployment")
plotPartial(partial_rf_Unemploy, main = "Partial Dependence Plot para Unemployment")

partial_rf_Fuel <- partial(rf_model, pred.var = "Fuel_Price")
plotPartial(partial_rf_Fuel, main = "Partial Dependence Plot para Fuel Price")

partial_rf_Temperature <- partial(rf_model, pred.var = "Temperature")
plotPartial(partial_rf_Temperature, main = "Partial Dependence Plot para Temperature")
importance <- varImp(rf_model, scale = FALSE)
```

Importance

```
rf_importance <- varImp(rf_model, scale = FALSE)
```

```
importance_df <- as.data.frame(rf_importance$importance)

importance_df$Feature <- rownames(importance_df)

# Normalización
importance_df$Scaled <- importance_df$Overall / sum(importance_df$Overall)

exclude_features <- c("sales_standar", grep("Store", importance_df$Feature, value = TRUE))
importance_df <- importance_df[!(importance_df$Feature %in% exclude_features),]

# Reescalar
importance_df$Scaled <- importance_df$Overall / sum(importance_df$Overall)

# Grafico
ggplot(importance_df, aes(x = reorder(Feature, Scaled), y = Scaled)) +
        geom_bar(stat = "identity") +
        coord_flip() +
        labs(x = "Característica", y = "Importancia Normalizada", title = "Importancia de las Características
        Normalizada") +
        theme_minimal()
```